Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

XYZ Sports Company – Customer Segmentation

Group 99

Guilherme Sá, number: 20230520

Sebastião Rosalino, number: 20230372

Zenan Chen, number: 20221390

January, 2024

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

IMAGE INDEX

Section 2 - Data Exploration:

Section 3.1 – 1st Outlier Removal:

Section 3.4 - Feature Engineering:

Section 3.5 - 2nd Outlier Removal:

Section 3.6 - Feature Selection:

Section 4.2 – Hierarchical Clustering:

Section 4.3 – K-Means:

Section 4.4 – Partitioning Around Medoids:

Section 4.5 – Self Organizing Maps:

Section 4.6 – Gaussian Mixture Models:

Section 4.7 – K-Prototyes:

Section 4.8 – Final Solution & Clustering by Perspective:

Section 5.1 – Outlier Reclassification

1. Introduction

In the dynamic landscape of the sports retail industry, effective customer segmentation stands as a key differentiator for businesses. This report presents an in-depth analysis and segmentation of the XYZ Sports Company's customer base, utilizing data from its Enterprise Resource Planning (ERP) system[1]. The project leverages advanced data mining methodologies to offer actionable insights for strategic decision-making and enhanced customer engagement. XYZ Sports Company, with its diverse clientele, faces the dual challenge and opportunity of understanding and addressing varied customer needs. Traditional demographic-based segmentation methods are inadequate for capturing the complexities of modern consumer behavior.

This project employs sophisticated data analysis techniques, transcending conventional approaches to reveal behavior-based customer segments. The essence of this project is transforming raw data into strategic insights. It involves a detailed data preparation phase to ensure the accuracy and reliability of the analysis. Advanced clustering techniques like K-Means and Gaussian Mixture Models are utilized to identify distinct customer groups. Each technique provides unique insights, and their comparative analysis ensures a thorough understanding of the customer base. Dimensionality reduction techniques, such as t-SNE, are employed for data visualization, facilitating the identification of patterns and relationships in high-dimensional data. This approach enhances both the analysis process and the presentation of findings, especially for stakeholders not versed in data science. Crucially, the project focuses on the interpretability of results. Understanding the defining characteristics of each customer segment is vital for practical applications, such as targeted marketing and personalized product recommendations.

This report demonstrates how data mining can be applied to uncover valuable insights within customer data. The findings are anticipated to provide XYZ Sports Company with a deeper understanding of their customers, supporting data-driven decisions that strengthen customer relationships and propel business growth.

2. Data Exploration

2.1 Data Exploration Overview

The exploration phase commenced with a comprehensive overview of the dataset. Initial observations revealed a wide range of features, offering insights into customer behaviors and attributes. However, to streamline the analysis, certain features were deemed non-contributory and subsequently removed. The **'ID'** feature, being merely an identifier, lacked analytical value for customer segmentation. Similarly, **'DanceActivities'** and **'NatureActivities'** were dropped due to their uniform values (all zeroes), indicating no customer engagement in these areas. This pruning of irrelevant features was crucial for focusing the analysis on meaningful data.

2.2 Feature Renaming Standardization

In pursuit of enhancing data readability, especially in visualizations, a systematic renaming of data frame columns from camelCase to snake_case was undertaken. This modification enhances the interpretation of the visualizations.

[1] P. Pinheiro and L. Cavique, “Regular sports services: Dataset of demographic, frequency and service level agreement,” *Data in Brief*, vol. 36, p. 107054, Jun. 2021, doi: 10.1016/j.dib.2021.107054.

2.3 Dropping and Transforming Features

The **'gender'** feature, initially encoded with textual labels ('Female' and 'Male'), was transformed into a binary format. This conversion simplifies future data processing and allows for a more straightforward approach, where binary features are more efficiently handled in numerical format. ‘Male’ is now represented with 1 and ‘Female’ with 0. The gender distribution in the dataset is as follows:

- Approximately 59.77% of the individuals in the dataset are labeled as "Female."
- About 40.23% of the individuals in the dataset are labeled as "Male."

2.4 Correlation Analysis

A key part of the exploration involved examining correlations among metric features. A Pearson correlation matrix was constructed, revealing several significant relationships:

- Moderate Positive Correlations: Notably, **'lifetime_value'** and **'attended_classes'** showed a moderate positive correlation, implying that customers with higher lifetime values tend to attend more classes. Similarly, **'number_of_frequencies'** and **'number_of_renewals**', as well as **'allowed_weekly_visits_by_sla'** and **'allowed_number_of_visits_by_sla'**, exhibited moderate positive correlations, suggesting coherence in visit patterns and membership renewals.
- Strong Positive Correlations: A strong positive correlation between **'age'** and **'income'** indicated a likely increase in income with age. Additionally, a robust link was observed between **'lifetime_value'** and **'number_of_renewals'**, highlighting the importance of customer value in membership continuity.
- Absence of Negative Correlations: The dataset did not exhibit any strong or moderate negative correlations among its metric features, suggesting no significant inverse relationships.

The entire correlation matrix can be seen in the **Annex, Figure 2.1.**

2.5 Visualizing Binary Features

For binary features, count plots were employed to visualize their distributions. This analysis revealed balanced distributions in the features ‘**gender’**, **'water_activities'**, and **'fitness_activities'**. However, imbalances were evident in several features, with ‘**use_by_time’, ‘athletics_activities’, ‘team_actitivities’, ‘racket_activities’, 'combat_activities', ‘special_activities’, 'other_activities'** and ‘**has_references’** showing significant skewness**.** The **'dropout'** feature, representing customer enrollment status, displayed a moderate imbalance, indicating potential areas for customer retention strategies. These exploratory steps laid the foundation for deeper segment analysis, setting the stage for a nuanced understanding of XYZ Sports Company's customer base.

Every count plot can be seen in the **Annex, Figure 2.2.**

3. Data Preprocessing

3.1 First Outlier Removal

Metric Features

In the Data Preprocessing phase, the initial outlier removal involved a meticulous examination of metric features using boxplots and histograms, as detailed in **Annexes, Figures 3.1.1** and **3.1.2.** The focus was on removing extreme outliers by applying specific thresholds for each metric feature.

Observations were excluded if they exceeded any of these limits: **'lifetime_value'** above 6000, **'attended_classes'** over 500, **'allowed_number_of_visits_by_sla'** greater than 200, **'real_number_of_visits**' more than 70, and **'number_of_references'** exceeding 2. This selective removal, which accounted for approximately **<u>0.4%</u>** of the dataset, ensured a balance between preserving data integrity and eliminating highly divergent data points.

Post-removal Data Visualization

After outlier removal, the dataset's metric features were re-evaluated using boxplots and histograms, as shown in **Annexes, Figures 3.1.3** and **3.1.4.** This post-removal visualization confirmed a more balanced distribution, indicative of a dataset that better represents the typical customer, while retaining diversity. This refined dataset lays a solid foundation for accurate customer segmentation analysis, ensuring that insights and conclusions are both reliable and reflective of the actual customer population.

Binary Features

Binary features with disproportionate value distributions were identified. Those **<u>with a dominant binary value exceeding 94%</u>** were considered less informative and excluded from the analysis. This exclusion of the features **'use_by_time'**, **'athletics_activities'**, **'team_activities'**, **'racket_activities'**, **'special_activities'**, **'other_activities'** and **'has_references'** ensures a balanced set of features, essential for a representative segmentation. This method enhances the dataset's robustness, providing a more reliable foundation for segmentation analysis. Every count plot supporting this decision can be visualized in **Annex, Figure 3.1.5.**

Date Features

The key metric features of **'enrollment_duration'** and **'last_period_duration'** were derived to evaluate customer commitment. The feature **'date_last_visit'** was also converted to the number of days (taking Jan 1, 1970, as reference date) to track the recency of facility engagement. Despite identifying numerous outliers in **'enrollment_duration'**, they were retained to avoid significant data loss, as they were numerous and clustered together. The other date features, being outlier-free (considering the ± 1.5 * IQR threshold), were deemed fit for further analysis. The distributions of the date-related features can be seen in **Annexes, Figures 3.1.6** and **3.1.7.**

3.2 Addressing Data Inconsistencies

To address data inconsistencies in the dataset, two key issues were tackled:

- Continuous Memberships: Customers with active, continuous memberships (with a '**dropout'** of 0) showed a positive **'last_period_duration'** but a zero **'enrollment_duration'**. This anomaly was due to the ongoing nature of their memberships, lacking an **'enrollment_finish'** date. The solution involved updating the **'enrollment_finish'** date to the most recent **'last_period_finish'** date (Dec 31, 2019) for these customers. This recalibrated **'enrollment_duration'** to accurately represent the duration between **'enrollment_start'** and the updated **'enrollment_finish'**.
- Frequency vs. Visits Inconsistency: An illogical scenario was identified where **'number_of_frequencies'** (total visits since the first enrollment) was less than **'real_number_of_visits'** (visits in the last active period). Logically, **'number_of_frequencies'** should be equal to or greater than **'real_number_of_visits'**. This was addressed by equalizing **'number_of_frequencies'** to **'real_number_of_visits'** in affected records, maintaining logical coherence and accurately reflecting customer engagement.

3.3 Missing Data Handling

Metric Features

In addressing missing data within the metric features **'income,' 'number_of_frequencies,'** and **'allowed_weekly_visits_by_sla',** a two-step process was employed. Initially, data standardization was applied exclusively to metric features, ensuring their comparable scale (mean of 0 and standard deviation of 1). Subsequently, a K-Nearest Neighbors (KNN) Imputer, with 5 nearest neighbors, was utilized to impute missing values. This supervised technique restores metric feature integrity while preserving the original data's statistical distribution.

Binary Features

To handle missing data in binary features, a Logistic Regression was applied as a supervised imputation method. Key binary features with missing values were pinpointed (**'water_activities', 'fitness_activities', 'combat_activities')**, and the imputation model used the dataset's standardized metric features as inputs. The model was trained on data without missing values and then used to predict the missing binary data. This approach, detailed in **Annex, Figure 3.3.1**, ensured accurate imputation, leveraging correlations with metric features to restore missing binary data effectively.

3.4 Feature Engineering

Binary Feature

- The **'has_references'** binary feature, indicating the presence of family or friendship relationships through 'number_of_references' greater than zero, was found to be consistently univariate at class 0. This indicated a uniform lack of such relationships between customers. Due to its lack of variability and limited informational value, the decision was made to drop this feature. This removal aims to streamline the efficiency of the upcoming customer segmentation analysis. Proof of this is provided in **Annex, Figure 3.4.1**.

Categorical Feature

- **'age_groups'**: This new age segmentation feature categorizes customers into five groups: 0-19, 20-34, 35-48, 49-64, and 65+. Analysis shows a majority (53.92%) in the 20-34 age group, followed by 25.45% in the 0-19 bracket, indicating a young customer base. Other age groups include 35-48 (12.40%), 49-64 (5.91%), and 65+ (2.32%), highlighting a diverse age range. This demographic distribution is essential for developing targeted business strategies to cater to each age segment's needs. Proof of this is provided in **Annex, Figure 3.4.2.**

Metric Features

- **Weekly Average Frequency Calculation:** Computes the **'weekly_average_frequency'** feature. This is done by dividing the total number of frequencies by the enrollment duration, with the time frame considered in weeks.
- **Total Activity Engagement:** Aggregates various types of activities, namely water, fitness, and combat activities, into a single feature. The sum of these activities forms the **'number_of_activities'** feature, representing a user's total engagement in different activities.
- **Average Weekly Classes Attendance:** Calculates the **'attended_classes_weekly_average'**, which reflects the average number of classes attended by a user per week throughout their enrollment period.
- **Visits Ratio Feature:** Establishes a **'visits_ratio'** feature that compares the actual number of visits a user has made to the number of visits allowed as per the service level agreement (SLA). This ratio provides insight into how the user's actual visitation patterns align with the permitted standards.

The **'visits_ratio'** feature, representing the ratio of actual to allowed visits, showed values exceeding 1, suggesting data inconsistencies. These instances were normalized by setting **'visits_ratio'** to 1, based on the assumption that excessive visits might be due to special events or open days. This adjustment ensures that the data accurately reflects customers' adherence to their visit allowed quotas during their last registration period. Proof of this in the **Annex, Figures 3.4.3** and **3.4.4.**

3.5 Second Outlier Removal

This final outlier removal stage, occurring after all prior preprocessing activities, aims to ensure data integrity and bolster the effectiveness of the subsequent customer segmentation analysis.

Manual Removal

For each metric feature, outliers were manually removed based on pre-determined thresholds, based on **Annex, Figures 3.5.1** to **3.5.18.** A detailed table, which can be seen in **Annex, Figure 3.5.19**, outlines the specific thresholds applied, the number of records removed, and the percentage of data loss for each feature. This targeted removal, based on careful threshold selection, ensured the elimination of extreme data points while minimizing information loss.

Removal with DBSCAN

Following the initial manual removal of outliers, the DBSCAN algorithm was employed for a more refined approach. Using an Elbow Plot, which can be seen in **Annex, Figure 3.5.20**, parameters were optimized; the number of neighbors was set at 10 based on the plot's 'elbow' point, and the epsilon value was determined at **3.5**, marking the increase in the distance to the 10th nearest neighbors, which can be seen in **Annex, Figure 3.5.21.** This method effectively differentiated between dense cluster points and potential outliers. Resulting in the exclusion of 220 data points, the process achieved a final 3.8% reduction in the dataset, enhancing its quality for customer segmentation analysis.

Binary Features

Post-outlier removal, the analysis of binary features' count plots, which can be seen in **Annex, Figure 3.5.22** indicates a well-balanced class distribution, affirming the effectiveness of the outlier removal process. The **'combat_activities'** feature, with **its least represented class at 10.25%**, highlights this balance. This equilibrium across binary features marks a successful data preprocessing phase, setting the stage for a more robust and representative dataset, vital for the forthcoming customer segmentation analysis.

Categorical Features

Recognizing the potential significance of age as a crucial factor in customer segmentation, a meticulous examination of the age group distribution was undertaken, which can be closely seen in **Annex, Figure 3.5.23.**

- **20-34:** 7870 individuals
- **0-19:** 3669 individuals
- **35-48:** 1738 individuals
- **49-64:** 800 individuals
- **65+:** 293 individuals

The age group analysis reveals distinct numerical representations across various brackets. This diversity suggests unique characteristics and behaviors within each age group, affirming the decision to retain all age categories in the analysis. Maintaining this granularity enables a comprehensive exploration of customer dynamics across life stages. This strategic approach is key for developing targeted strategies that cater to the specific needs and behaviors of different age groups, ensuring an effective and nuanced understanding of customer preferences.

3.6 Feature Selection

The feature selection stage was integral in optimizing the clustering model for customer segmentation. This critical step employed various methods to evaluate and select the most relevant features across metric, binary, and categorical types, ensuring a comprehensive and effective segmentation model.

Feature Selection on Metric Features

- **Correlation Matrix Approach:** The features **'age'** and **'income'** showed redundancy, leading to the removal of **'income'** due to its original higher missing data proportion **(3.30%)** and similar correlation scores with other variables. Strong correlations were observed between **'number_of_renewals'** and **'enrollment_duration',** and **'lifetime_value'** and **'enrollment_duration'.** These logical correlations, in line with business insights, prompted the use of Self Organizing Maps (SOM) for a nuanced feature analysis. SOMs helped refine the feature set, balancing logical relationships and data integrity for a more insightful clustering solution. The full correlation matrix can be seen in **Annex, Figure 3.6.1.**
- **Self-Organizing Maps (SOM) Approach:** In refining the Feature Selection using Self-Organizing Maps, **'age'** was retained over **'income'** for its relevance and lack of missing data. Features showing minimal variation, like **'days_without_frequency'** and **'number_of_frequencies'**, were omitted due to their limited differentiation potential. **'attended_classes'** was replaced with **'attended_classes_weekly_average'** for more detailed weekly engagement insights. **'allowed_number_of_visits_by_sla'** was dropped in favor of **'allowed_weekly_visits_by_sla'** for more detailed engagement insights. Despite low variability, **'number_of_activities'** and **'visits_ratio'** were kept for their unique insights into customer activity diversity and visitation patterns, essential for effective customer segmentation. See the full SOM component planes here in **Annex, Figure 3.6.2.**

Feature Selection on Binary Features

- **Phi Coefficient Heatmap:** The Phi Coefficient Heatmap was used to analyze binary features, whose detailed explanation and obtained results are in **Annex 3.6.3.** The analysis identified **'gender'** and **'dropout'** as potential candidates for removal due to their minimal and modest associations with other binary features, respectively[2].
- **Boxplots & Violin plots Analysis:** Further insights on the variance of the binary features along the metric features were gained through Boxplots and Violin plots. The feature **'gender'** showed insignificant variation among the metric features, while the remining features showed relevant variability among the metric features. The Boxplots and Violin plots can be seen in **Annex, Figure 3.6.4.**
- **Point-Biserial Correlation Matrix:** This analysis, whose detailed explanation and obtained result are **Annex 3.6.5**, affirmed the insignificance of **'age'** and **'combat_activities'**, as they showed negligible correlations with metric features[3].

A summary table encapsulates the decisions made across different analyses for binary features:

[2] "Phi Coefficient (Mean Square Contingency Coefficient) - Statistics How To." Accessed: Jan. 06, 2024. [Online]. Available: https://www.statisticshowto.com/phi-coefficient-mean-square-contingency-coefficient/

[3] "t-Test, Chi-Square, ANOVA, Regression, Correlation..." Accessed: Jan. 06, 2024. [Online]. Available: https://datatab.net/tutorial/point-biserial-correlation

Feature	Phi Coefficient Heatmap	Boxplots & Violin plots	Point-Biserial Correlation Matrix
Gender	Drop	Drop	Drop
Water Activities	Keep	Keep	Keep
Fitness Activities	Keep	Keep	Keep
Combat Activities	Drop	Keep	Drop
Dropout	Drop	Keep	Keep

As a result, **'gender'** and **'combat_activities'** were dropped due to most of the feature selection methods supporting this decision.

Feature Selection on Categorical Feature

The feature **'age'** was chosen for the final clustering solution due to its more impactful role in cluster formation, compared to **'age_groups'**, which may be utilized later for enhancing cluster interpretation.

4. Clustering

To find the ideal number of clusters, various techniques and evaluation scores were used. These include R^2 (indicating variance explained by clusters), Silhouette Score (measuring cluster separation, with values near +1 signifying clear clusters), Calinski-Harabasz Index (assessing cluster density and separation), and Inertia (reflecting internal cluster variance). For the detailed definition of the Calinski-Harabasz Index consult **Annex 4.1.** Only solutions with 3 or more clusters were considered, aligning with the business needs of XYZ Sports Company, as fewer clusters are deemed insufficient for capturing customer diversity and engagement patterns.

4.1 Hierarchical Clustering

In the Hierarchical Agglomerative Clustering analysis using various linkages and a range of 2 to 10 clusters, the optimal number was determined as 3 clusters. This decision was based on the significant peak in the Calinski-Harabasz index at 3 clusters for the 'ward' linkage, indicating well-defined clusters. Although the R^2 score improved with more clusters, the decline in Silhouette and Calinski-Harabasz scores beyond 3 clusters suggested that additional clusters did not enhance segmentation quality. Therefore, 3 clusters were selected as the best fit for the company's customer segmentation analysis. Detailed plots of these evaluation scores are available in **Annex, Figure 4.2.**

4.2 K-Means

For the K-Means clustering with n_init=500 and init='k-means++', the optimal number of clusters was determined as 3, based on evaluation metrics. Despite the Calinski-Harabasz Score peaking at 2 clusters, a minimum of 3 clusters was chosen to better align with the business context, offering a more nuanced segmentation. This decision is supported by the continuous increase in R^2 Scores, decline in Inertia and a high Silhouette, suggesting improved homogeneity and tighter clusters with more clusters. Detailed plots of these evaluation scores are available in **Annex, Figure 4.3.**

4.3 Partitioning Around Medoids

Partitioning Around Medoids (PAM), a clustering algorithm similar to K-Means, uses the most centrally located object in a cluster, the medoid, instead of the mean[4]. It's more resilient to noise and outliers

[4] M. Botyarov and E. E. Miller, "Partitioning around medoids as a systematic approach to generative design solution space reduction," *Results in Engineering*, vol. 15, p. 100544, Sep. 2022, doi: 10.1016/j.rineng.2022.100544.

since the medoid is an actual data point. The PAM algorithm iteratively searches for medoids that minimize dissimilarities within a cluster. The optimal number of clusters for this dataset is identified as 4, based on the Silhouette score, Calinski-Harabasz score, and Inertia, suggesting this number offers the best balance of cohesion, separation, and cluster compactness. Detailed plots of these evaluation scores are available in **Annex, Figure 4.4.**

4.4 Self-Organizing Maps

The Hierarchical Agglomerative and K-Means clustering applied to a 50x50 Self-Organizing Map grid successfully segmented the data into three distinct clusters. These methods demonstrated clear and homogenous cluster distributions with minimal overlap, as illustrated in **Annex, Figures 4.5.1** and **4.5.2**. The results indicate a robust and reliable clustering solution, with well-defined, compact clusters. The absence of stray units near other clusters' centroids in both methods further affirms the stability and effectiveness of these clustering techniques in capturing intrinsic customer segments in the dataset.

4.5 Gaussian Mixture Models

The evaluation of Gaussian Mixture Models clustering using BIC and AIC suggests that model improvement plateaus beyond 7 clusters. The Silhouette Score, peaking at 2 clusters, indicates optimal cluster separation at this level but shows inconsistency with more clusters. The Calinski-Harabasz Score decreases as clusters increase, suggesting less distinct clusters. Conversely, the R^2 Score consistently improves with more clusters. Considering XYZ Sports Company's business context, a 3-cluster solution is chosen for its balance of complexity, interpretability, and strategic alignment with customer segmentation objectives. Further details and plots are available in **Annex, Figure 4.6.1**.

4.6 K Prototypes

The K-Prototypes clustering combines K-Means and K-Modes, making it suitable for clustering mixed data types, including both numerical and categorical (even binary) data. It's particularly effective in scenarios where datasets contain binary features considered as categorical. The algorithm partitions the data into clusters by minimizing the dissimilarity between the categorical features and the prototypes of the clusters[5]. The 'Cao' method initialized centroids efficiently, balancing computational intensity with n_init set to 10. Optimal clustering favored 3 clusters, based on: strong Silhouette Scores at 3 clusters indicating good internal consistency and separation; high Calinski-Harabasz Scores for 3 clusters suggesting dense, well-defined clusters; and the Cost metric, balancing model complexity and clustering quality. Detailed plots of these evaluation scores are available in **Annex, Figure 4.7.1.**

4.7 Final Solution & Clustering by Perspective

The analysis of the several clustering methods leaves the resulting optimal values for each detailed in **Annex, Figure 4.8.**

The K-Means clustering method, **with a 3-cluster solution** for each customer perspective (Customer Value and Commitment & Customer Activity and Engagement), was selected for segmenting XYZ Sports Company's customer base due to its balance of segment diversity and model simplicity. Key reasons for this choice include its effective centroid initialization (k-means++), stability and reliability

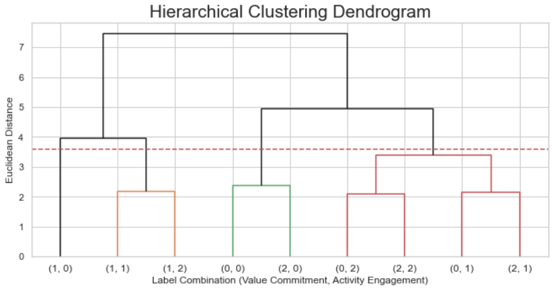
[5] Y. Reddy, “K-means, kmodes, and k-prototype,” Medium. Accessed: Jan. 06, 2024. [Online]. Available: https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669

(n_init=500) and robust clustering metrics **(Silhouette score ~0.26, Calinski-Harabasz score 3127.61, R^2 ~0.30)**, and computational efficiency.

Now onto the merging of the 9 created clusters, the below dendrogram offers a detailed illustration of the hierarchical agglomerative process applied to the K-Means-defined clusters. Each join in the dendrogram signifies the fusion of clusters, with the y-axis indicating the Euclidean distance, the greater the height of the join, the greater the dissimilarity between the merging clusters. **The chosen y-threshold for merging, set at 3.7 and highlighted by the dotted line**, is a deliberate decision to distinguish between clusters. This threshold ensures a clear demarcation, allowing for the identification of clusters with internal homogeneity while maximizing differentiation between them. Clusters merging below this threshold, such as those labeled (Value Commitment=1, Activity Engagement=1) and (Value Commitment=1, Activity Engagement=2), display strong within-group similarity, suggesting cohesive cluster characteristics. Conversely, the clusters (Value Commitment=0, Activity Engagement=0) and (Value Commitment=0, Activity Engagement=1), which merge above the threshold, are recognized as distinct entities. The rationale behind the 3.7 threshold is multifaceted:

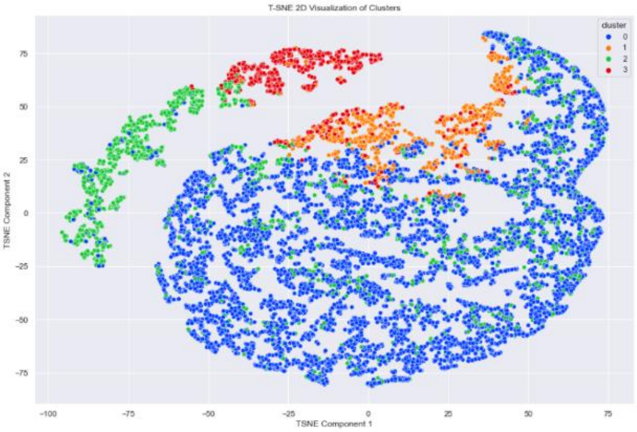
- Optimal Cluster Count: The threshold yields 4 discernible clusters, a number that balances complexity with the necessity of a meaningful customer segmentation.
- Cluster Separation: The selected distance indicates a natural partition within the data, where an increase in distance would be required to merge clusters, highlighting intrinsic separations.
- Business Context: For the XYZ Sports Company, 4 clusters afford a granular, yet manageable, understanding of customer demographics and engagement levels, aligning with strategic business applications, and allowing for the design of marketing strategies for each of the found clusters.

In essence, the dendrogram and threshold underscore the reliability of the clustering solution, capturing the diversity of the dataset in a manner conducive to targeted marketing and customer engagement strategies.



Hierarchical Clustering Dendrogram for Perspective Merging

4.8 Final Solution Visualization

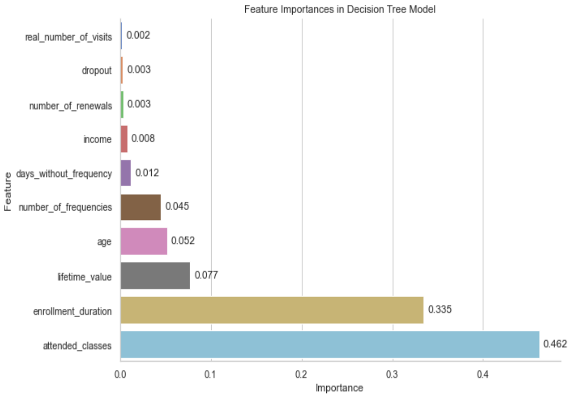


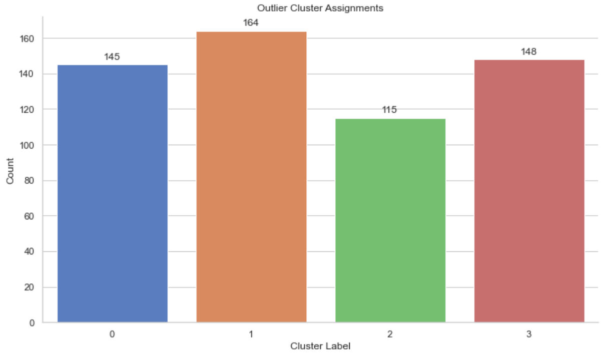
t-SNE 2D Visualization of the Final Clusters

The t-SNE 2D visualization of clusters offers an illustrative representation of the inherent groupings, providing valuable insights into the clustering distribution. It is evident that there is a clear delineation of clusters, each occupying distinct regions in the 2D space, which suggests that the clustering algorithm has effectively identified separable groups. Cluster 0 (blue), appears to be the most dispersed, indicating a greater within-cluster variance of behaviors within this group. This could represent a segment with diverse characteristics but enough commonalities to be grouped together. Cluster 1 (orange) shows moderate dispersion and is well separated from Clusters 2 and 3 but shares a boundary with Cluster 0. This might highlight a transitional group sharing traits with both the most and least cohesive clusters. Cluster 2 (green) is somewhat interspersed with Cluster 0 but still maintains a degree of separation. This suggests overlapping characteristics with Cluster 0 but with distinct differences that justify a separate grouping. Cluster 3 (red) is more tightly grouped and distinct from the others, implying a high degree of similarity among its observations. Such cohesion within a cluster often indicates strong defining features that are consistent across members of this group. The visualization underpins the clustering solution's robustness, with each cluster's concentration and separation providing a visual affirmation of the underlying patterns the algorithm has captured. It also suggests potential areas for further investigation, particularly where clusters meet, to understand the nuanced differences or similarities between those observations.

5. Outlier Reclassification

After obtaining the final clustering solution, a subsequent analysis was conducted to assess the results, including feature importance evaluation and reclassification of initially excluded outliers. A Decision Tree model, employing a grid search technique for hyperparameter optimization (criteria: 'gini', max_depth: 10, min_samples_leaf: 1, min_samples_split: 10), was used to predict cluster labels from metric features in the clustering. The data split was 80% training and 20% testing. The model showed a **94.02%** accuracy in predicting test set labels. Feature importance analysis from the Decision Tree identified key predictors for cluster labels: **Attended Classes (46.20%)** as the primary factor, **Enrollment Duration (33.50%)**, **Lifetime Value (7.70%)**, **Age (5.20%)**, and **Number of Frequencies (4.50%)**. Lesser impact predictors included Income, Days Without Frequency, and Number of Renewals, each under 1%. For outlier reclassification, identical preprocessing steps as the main analysis were applied, including feature selection, engineering, and missing value imputation, using the same scalers to maintain consistency and prevent data leakage. Post-processing, the outliers showed an even distribution across clusters.





Feature Importance in the Decision Tree Model

Outlier Cluster Assignments

The Decision Tree constructed with the tuned hyperparameters for reclassifying observations identified as outliers during the Preprocessing phase can be accessed in the ‘dt_reclassification_outliers.svg’ file on the sent zip. This model played a pivotal role in ensuring the accurate inclusion of these outliers into the final cluster framework. To visualize the comprehensive results, including both outliers and inliers, the final Cluster Assignment distribution is presented in

Annex, Figure 5.1.1. This representation offers a detailed view of how each observation, previously categorized as an outlier or inlier, has been systematically integrated into the final clustering solution.

6. Conclusion

To sum up the results, below are shown the answers to the initial expected outcomes as the main conclusions of this project.

a) Identified Customer Segments

The analysis has culminated in the identification of four distinct customer segments within the XYZ Sports Club membership:

- **Cluster 0**: Young adults (<30 years old), medium-high income, infrequent visits, low spending, and high dropout rate.
- **Cluster 1**: Mid-aged adults (>30 years old), highest income, frequent visits, high spending, active and loyal membership.
- **Cluster 2**: Adolescents (<20 years old), moderate income, low engagement, short enrollment duration.
- **Cluster 3**: Kids with balanced engagement, high frequency, and the highest spending.

b) Justification for the Number of Clusters

The decision to delineate four clusters was driven by the diverse demographic and behavioral patterns observed within the data. This granularity allows for the tailoring of marketing and operational strategies to address the unique preferences and tendencies of each segment effectively.

c) Explanation of Clusters

- **Cluster 0** represents a segment that, despite their potential, exhibits low engagement, necessitating strategies to boost visit frequency and reduce churn.
- **Cluster 1** comprises the most valuable members, guaranteeing premium services and rewards.
- **Cluster 2** includes the youngest members, where engagement can be fostered through targeted programs.
- **Cluster 3** represents the prospects, the kids, indicating a need for family-oriented offerings.

d) Business Applications and Marketing Approaches

- **Cluster 0**: Target with budget-friendly offers and loyalty discounts to increase engagement and reduce dropout rates. Digital marketing campaigns focusing on the cost-effectiveness of memberships may appeal to this segment.
- **Cluster 1**: Offer premium membership plans with exclusive benefits to capitalize on their higher income and loyalty. Personalized communication highlighting advanced facilities and loyalty rewards could foster long-term retention.
- **Cluster 2**: Engage with community and school programs to increase their attachment and regularity. Promotions of group activities and family memberships might be effective.
- **Cluster 3:** Introduce programs and events for skill development and fun activities that can attract kids and their parents. A family-centric marketing approach with a focus on the long-term health and educational benefits could be successful.

For each cluster, it is important to consider tailored communication strategies and personalized experiences to meet their specific needs and preferences, thereby enhancing customer satisfaction and retention. By addressing the specific needs of each segment, XYZ Sports Club can optimize its service offerings, bolster member satisfaction, and drive growth.

ANNEXES

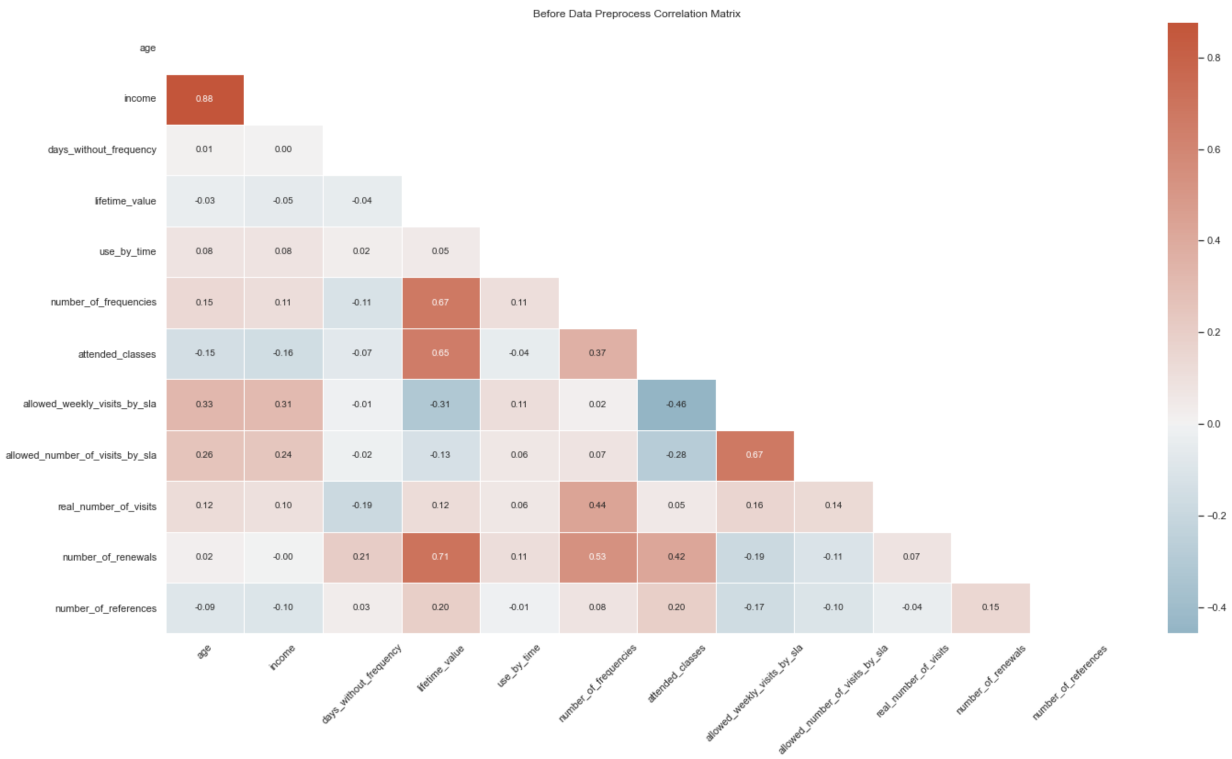


Figure 2. 1 – Before Data Preprocess Correlation Matrix / **Annex, Figure 2.1.**

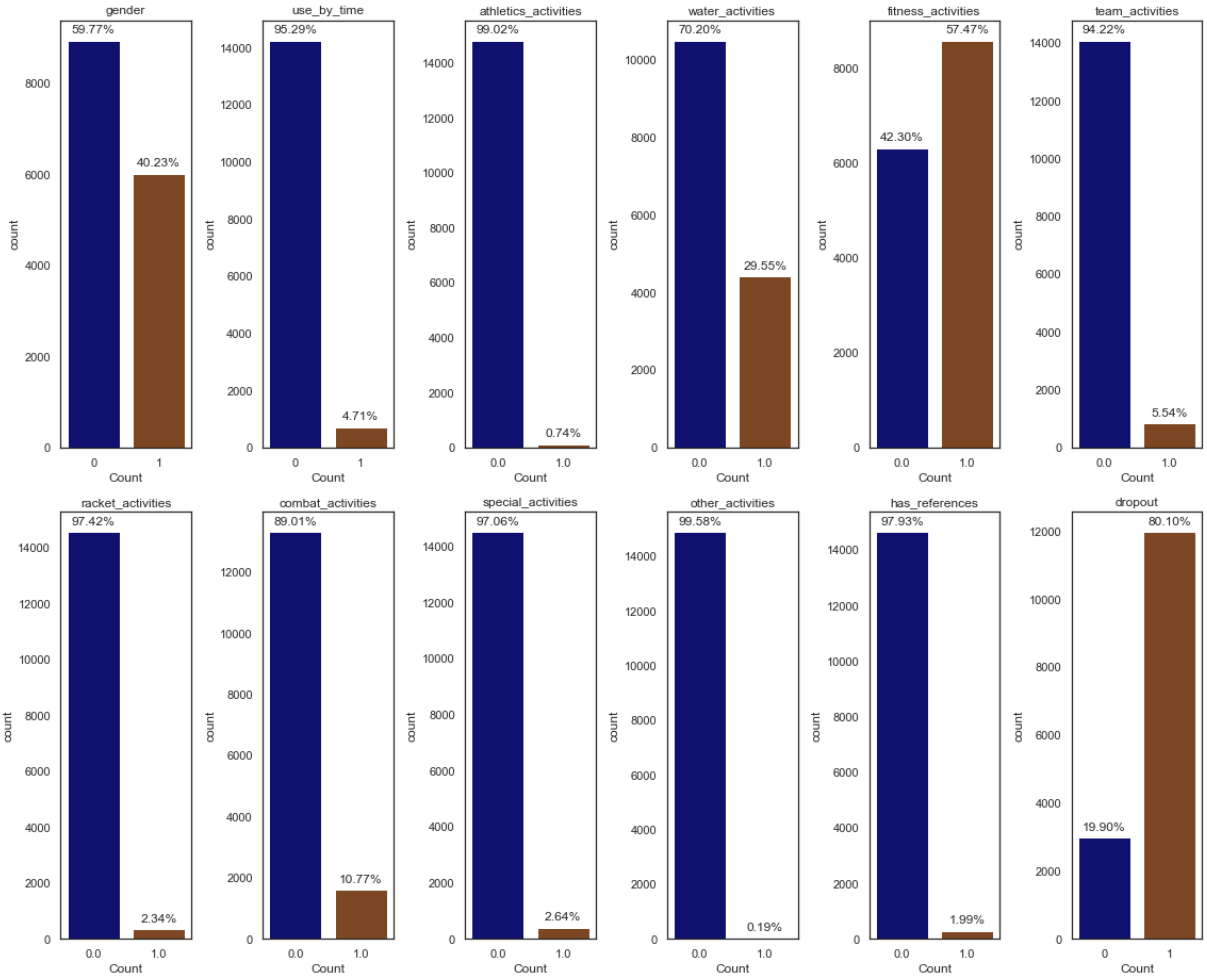


Figure 2. 2 – Bar plots of the Binary Features / **Annex, Figure 2.2.**

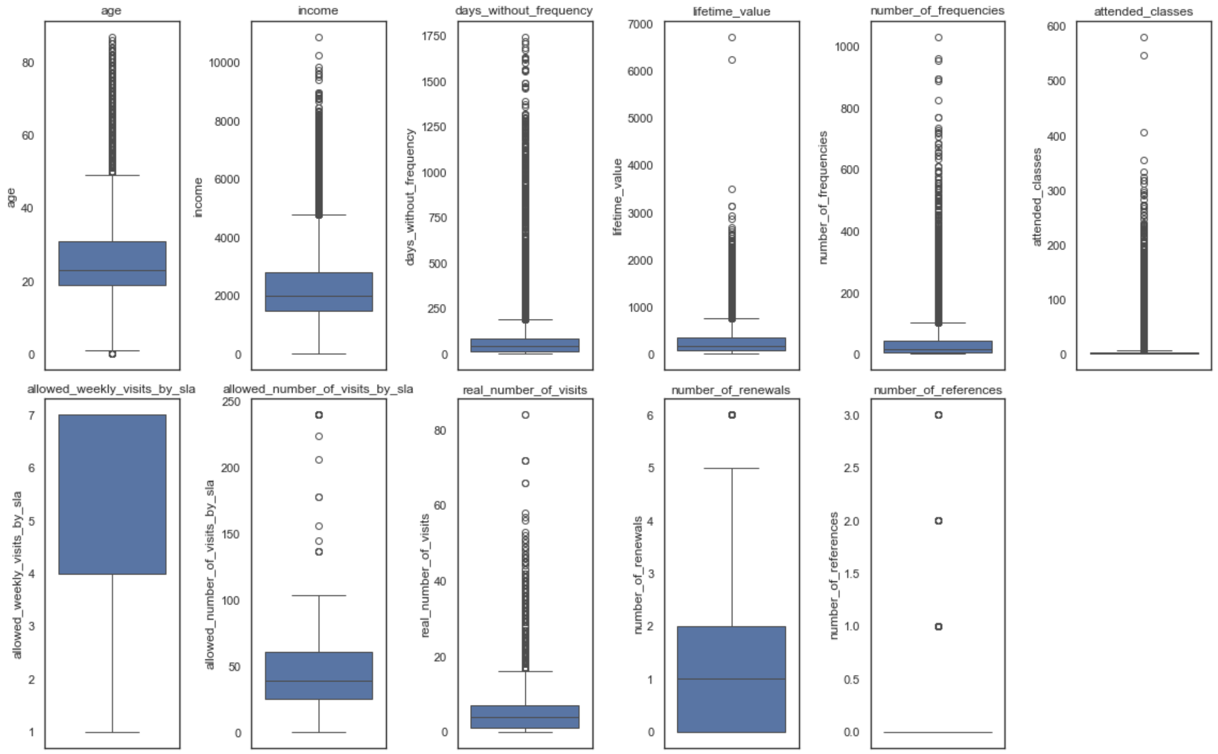


Figure 3.1. 1 – Boxplots of the Metric features before 1st Outlier Removal / **Annexes, Figures 3.1.1**

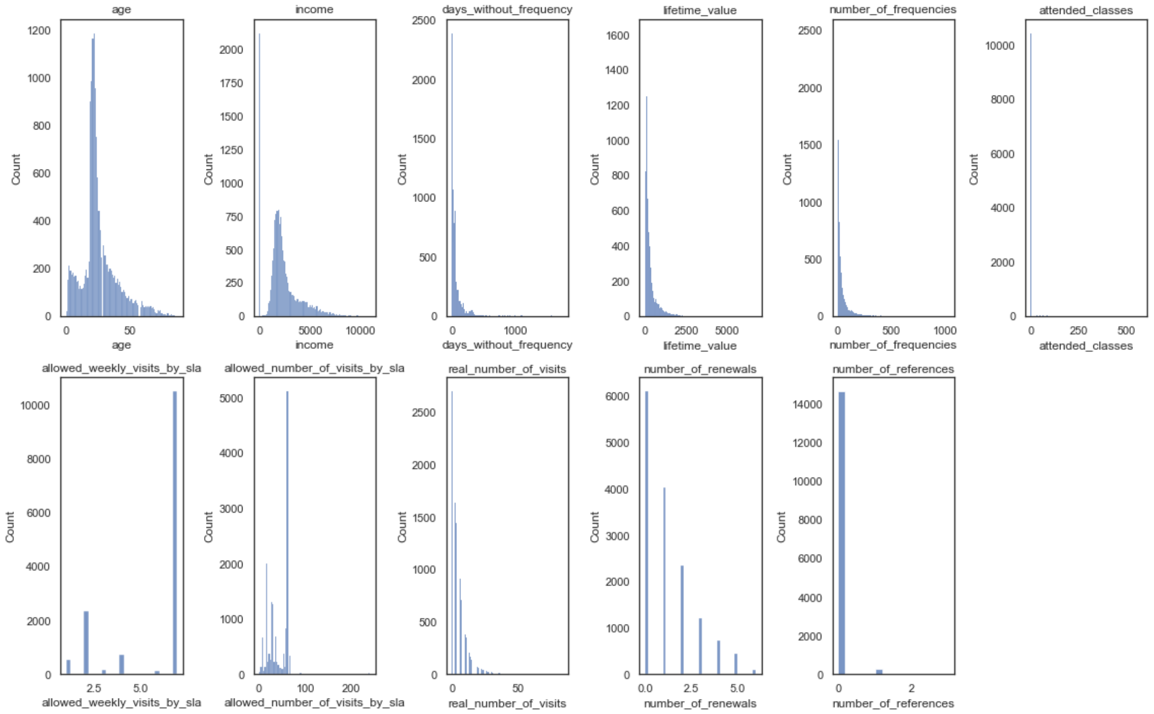


Figure 3.1. 2 – Histograms of the Metric features before the 1st Outlier Removal / **Annexes, Figures 3.1.1**

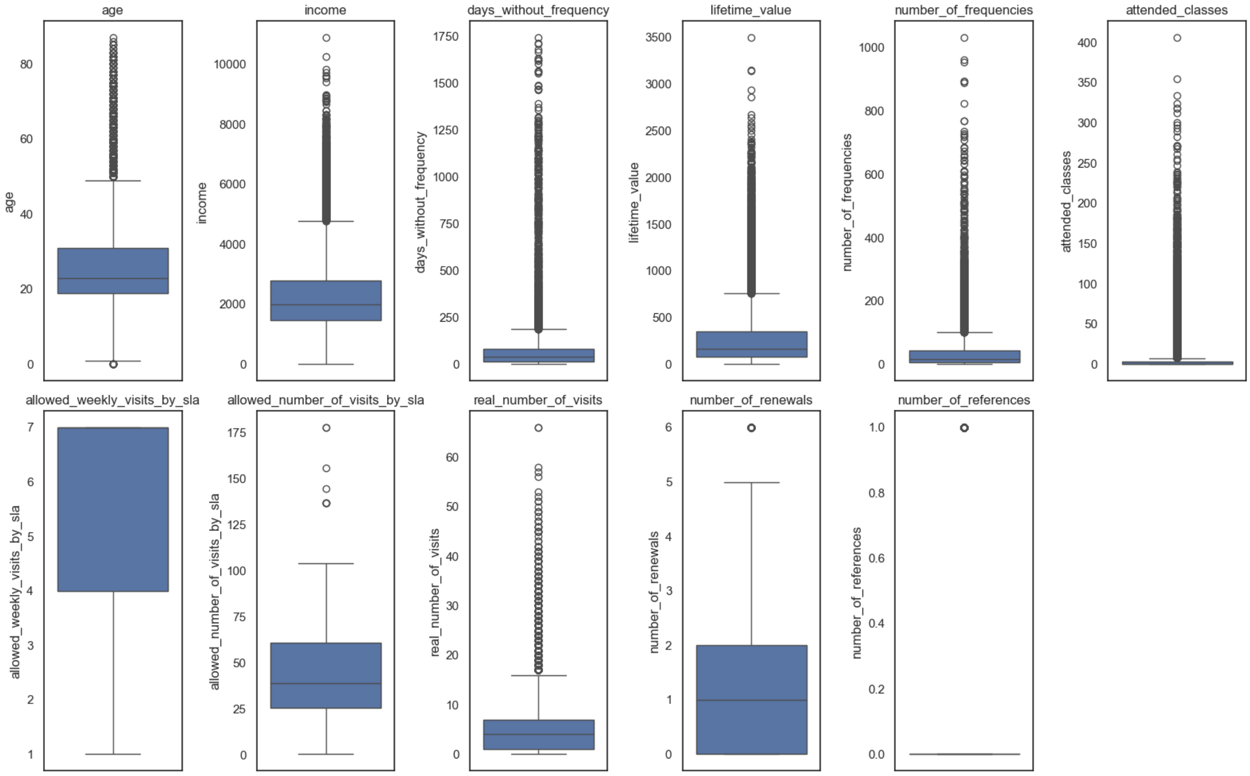


Figure 3.1. 3 – Boxplots of the Metric features after 1st Outlier Removal / **Annexes, Figures 3.1.3**

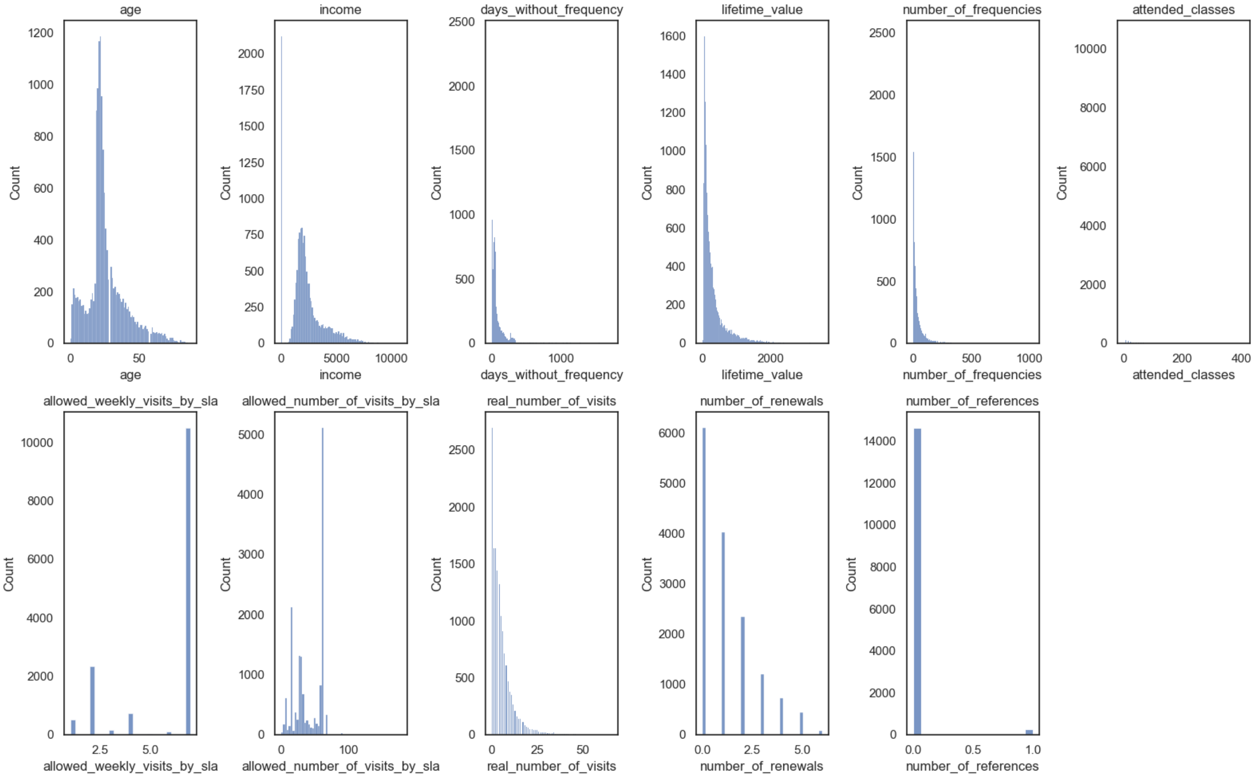


Figure 3.1. 4 – Histograms of the Metric features after the 1st Outlier Removal / **Annex 3.1.4.**

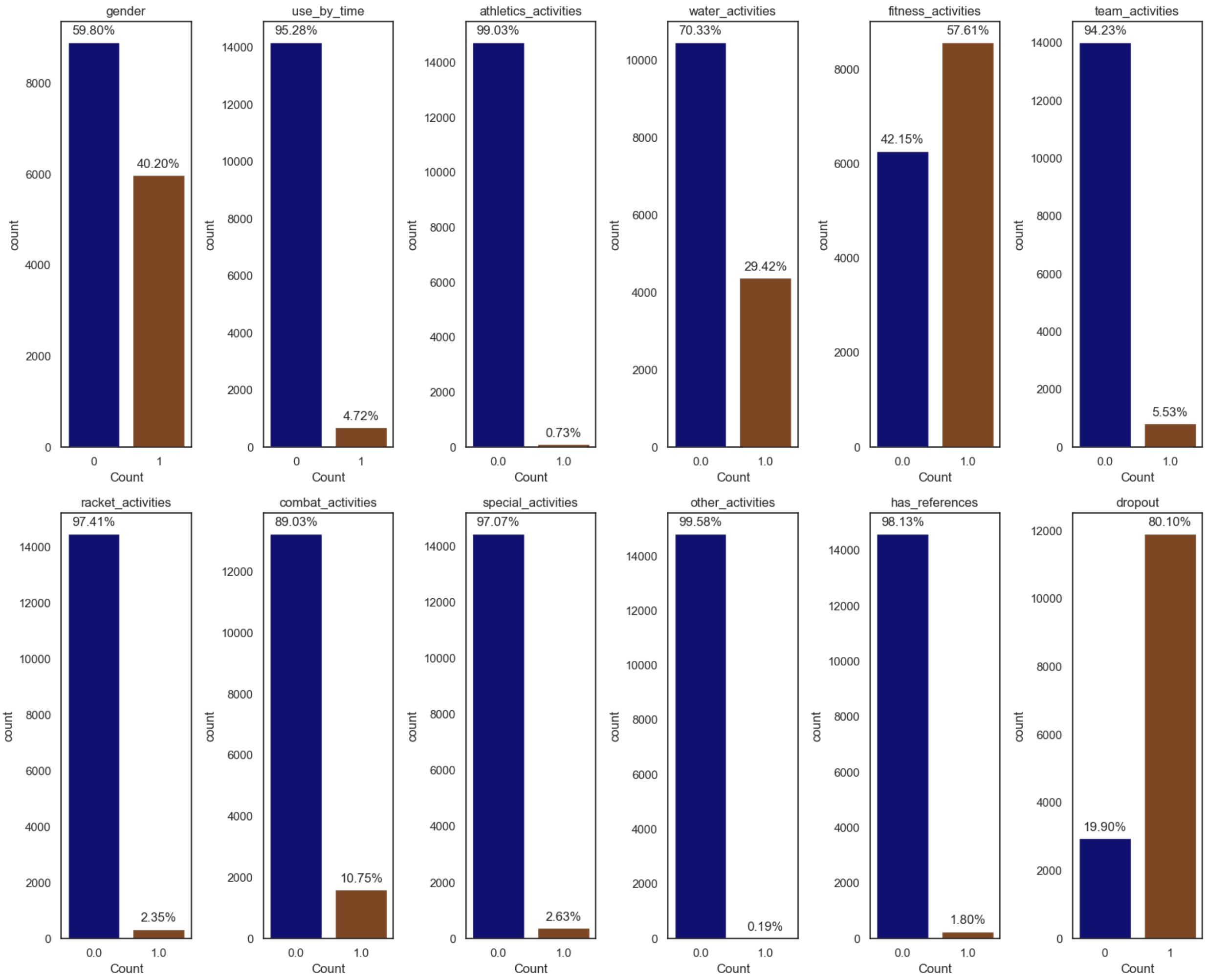


Figure 3.1. 5 – Bar plots of the Binary Features for the 1st Outlier Removal / **Annex, Figure 3.1.5.**

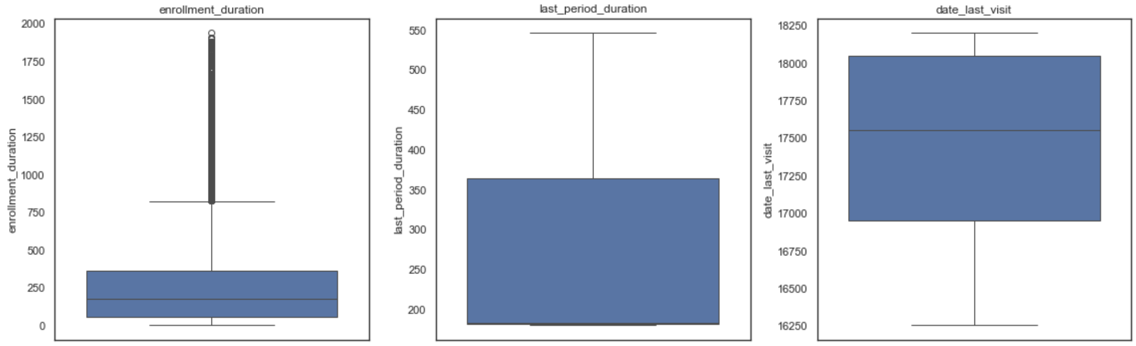


Figure 3.1. 6 – Boxplots of the Date Features / **Annexes, Figures 3.1.6**

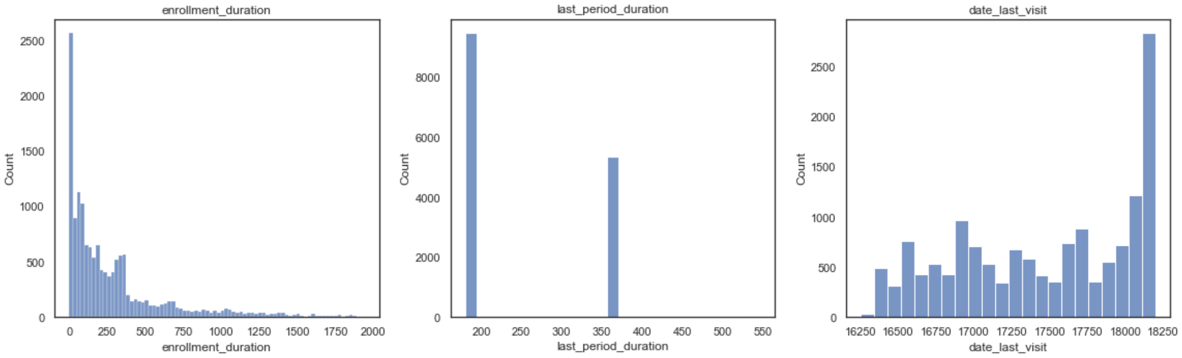


Figure 3.1. 7 – Histograms of the Date Features / **Annex 3.1.7.**

Binary Features Missing Data Handling Imputation Approach with Logistic Regression Explanation

Step 1 - Input Feature Definition:

- In the initial step, the binary feature necessitating imputation, referred to as the "target_feature," is identified.
- The selection of input features for the Logistic Regression model is made. These input features encompass all metric features, excluding the target binary feature. This step ensures that the imputation model exclusively relies on the available standardized metric features of the dataset.

Step 2 - Data Preparation:

Two distinct datasets are constructed:

- 'training_data' comprises solely those rows that contain non-missing values for the target binary attribute. This dataset is exclusively employed for the training of the imputation model.
- 'missing_data' comprises rows that contain missing values for the target binary attribute. These rows represent the dataset requiring imputation.

Step 3 - Model Selection:

- Initialization of a Logistic Regression model, a suitable choice for imputing binary features in a supervised learning context.

Step 4 - Model Training:

- The model training process commences with the preparation of training data. This includes input features ('X_train') and the target binary feature ('y_train').
- The Logistic Regression model is trained on the training dataset. Through this training, the model becomes capable of predicting the target binary attribute based on the provided input features.

Step 5 - Model Prediction:

- The data containing missing values for the input features ('X_missing') is prepared for the subsequent model prediction stage.
- The trained Logistic Regression model is effectively deployed to predict the missing binary values. It generates predictions for the rows housing missing data within the 'missing_data' dataset.

Step 6 - Imputation:

- In the concluding step, the missing binary feature values present within the original dataset ('no_outlier_no_missing_df') are substituted with the predicted values ('imputed_values'). This finalizes the imputation of missing binary data using the Logistic Regression model. This approach ensures the supervised imputation of binary features, incorporating information from other dataset attributes and maintaining dataset structural integrity. The process facilitates dataset completion while preserving information reliability, preparing the data for subsequent analyses and modeling.

Explanation 3.3.1 - Binary Features Missing Data Handling with Logistic Regression Explanation / **Annex, Figure 3.3.1**

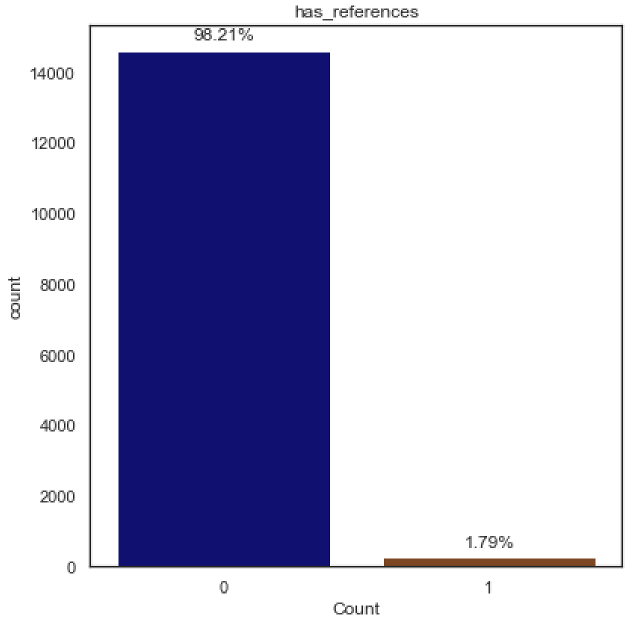


Figure 3.4. 1 – Bar plot of 'has_references' / **Annex, Figure 3.4.1**.

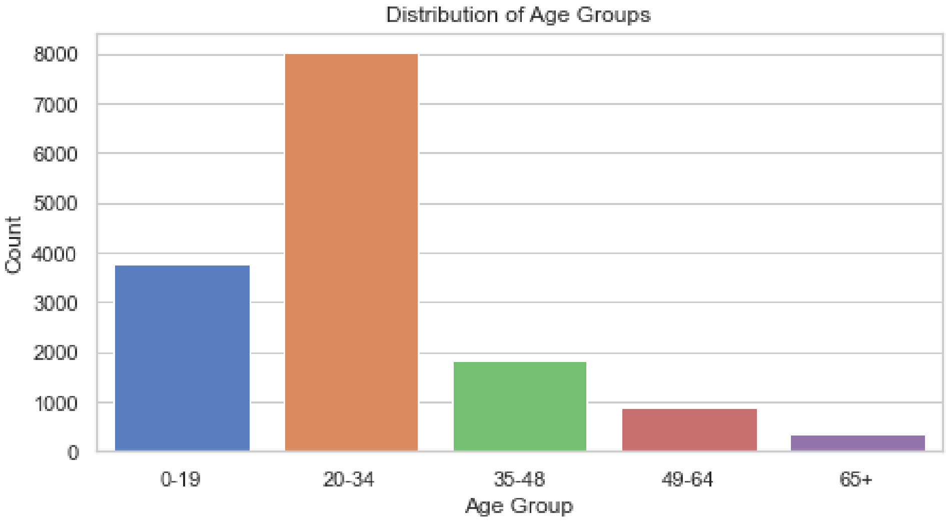


Figure 3.4. 2 – Bar plot of 'age_groups' / **Annex, Figure 3.4.2.**

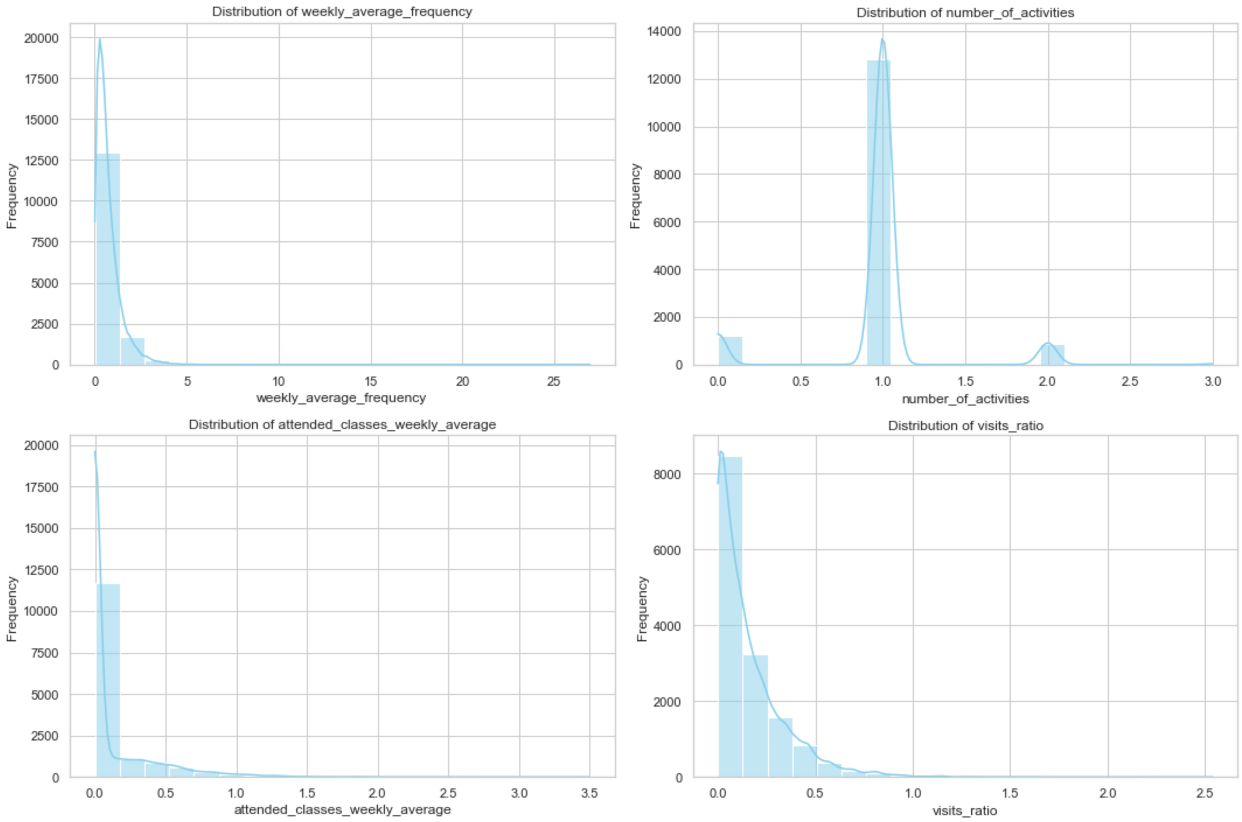


Figure 3.4. 3 – Histograms of the new metric features / **Annex, Figures 3.4.3**

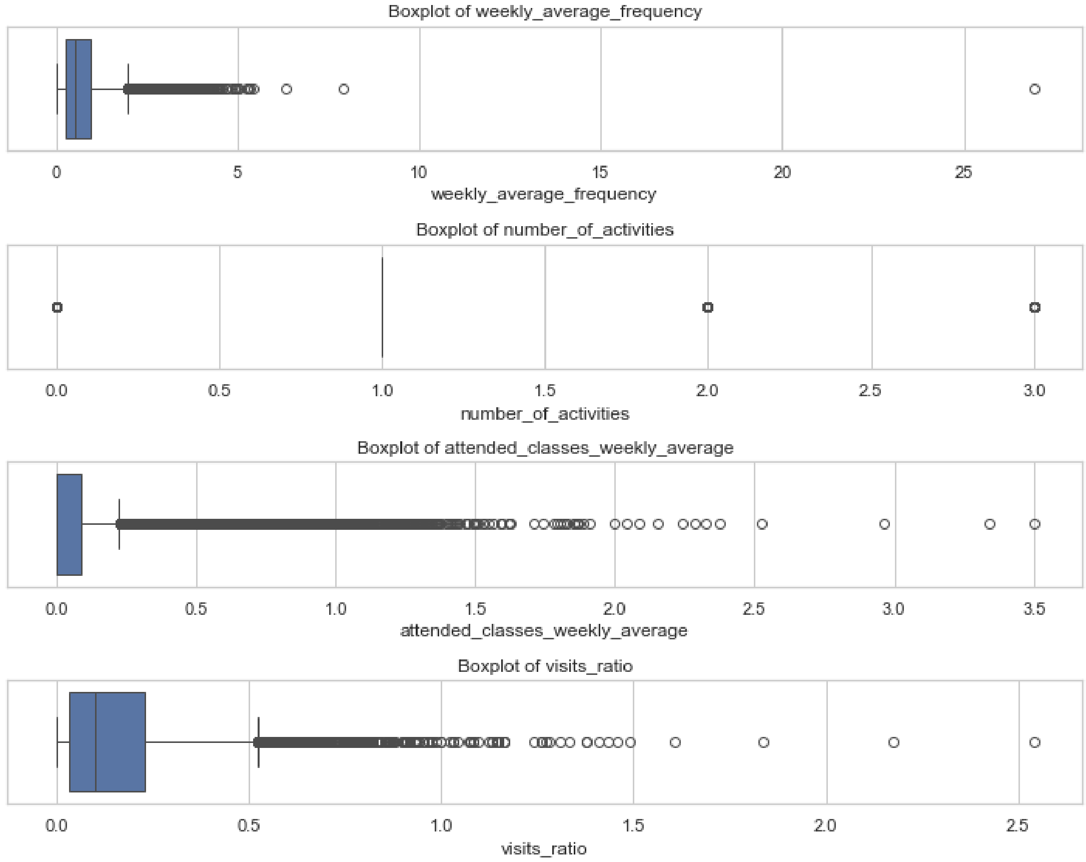


Figure 3.4. 4 – Boxplots of new metric features / **Annex 3.4.4.**

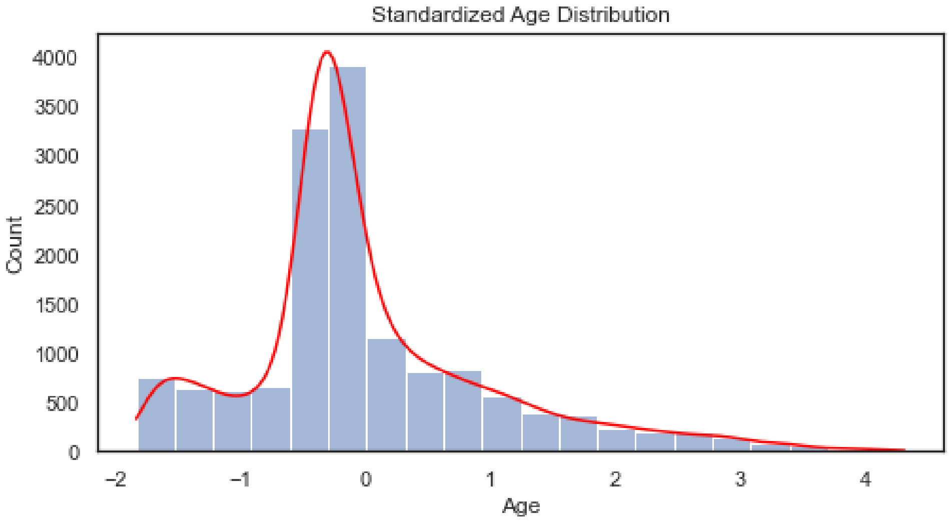


Figure 3.5. 1 – Histogram of Standardized Age Distribution / **Annex, Figures 3.5.1**

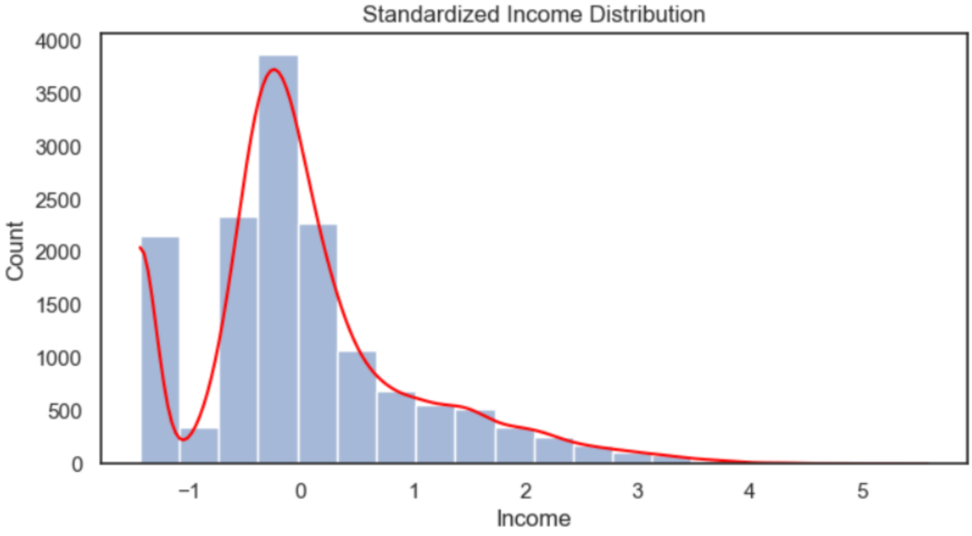


Figure 3.5. 2 – Histogram of Standardized Income Distribution / **Annex, Figures 3.5.1**

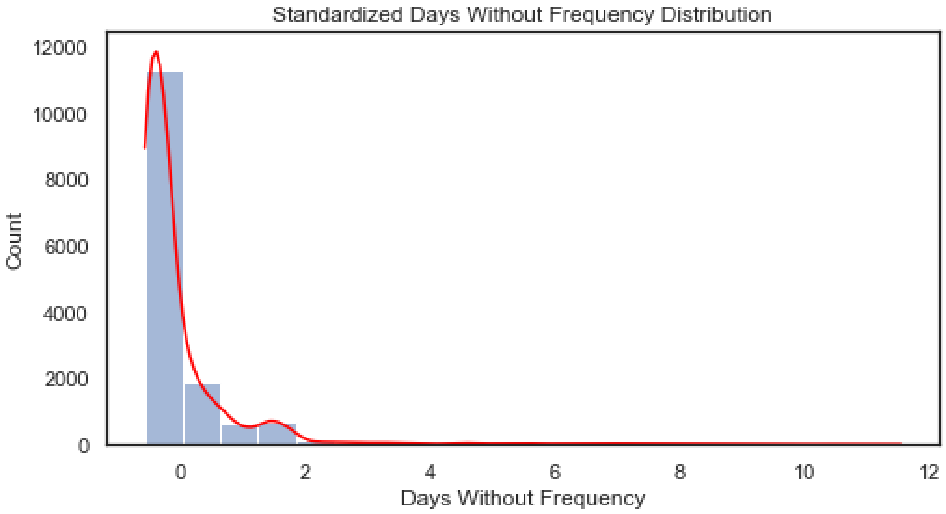


Figure 3.5. 3 – Histogram of Standardized Days Without Frequency Distribution / **Annex, Figures 3.5.1**

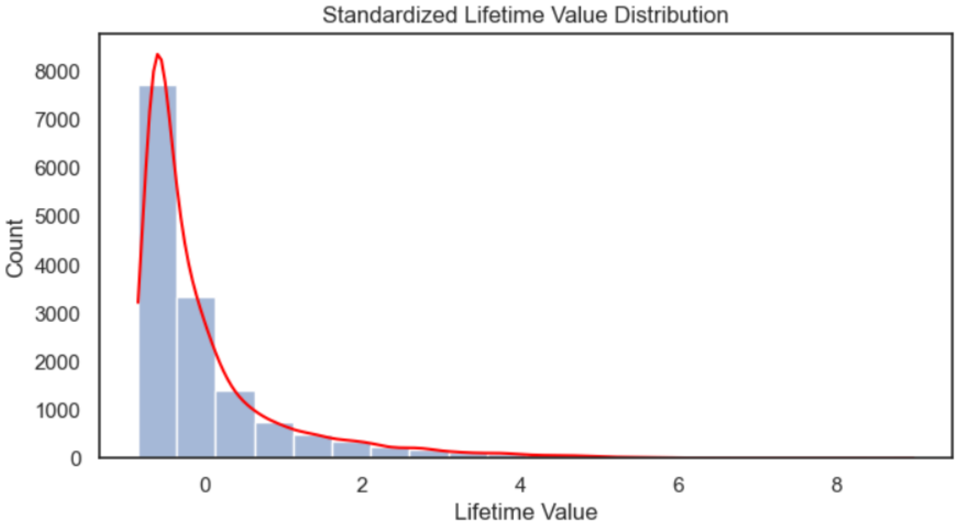


Figure 3.5. 4 – Histogram of Standardized Lifetime Value Distribution / **Annex, Figures 3.5.1**

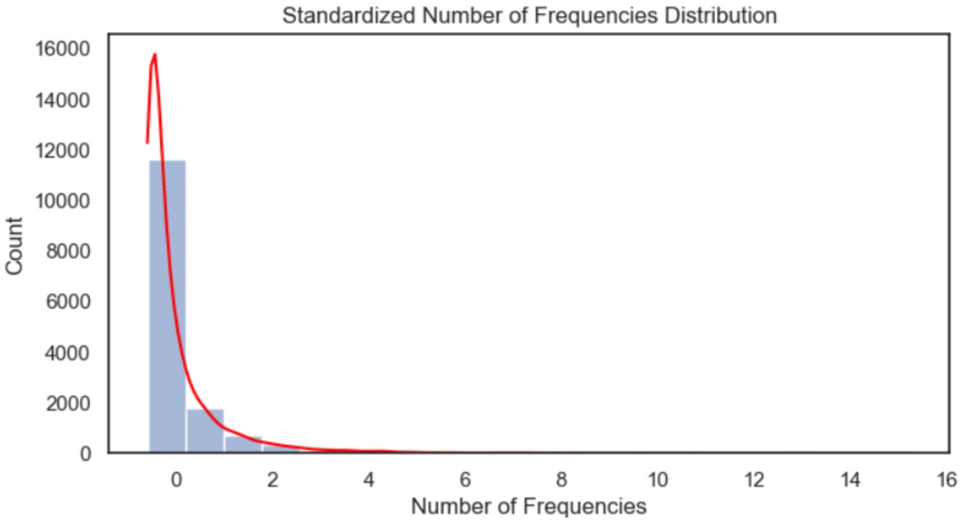


Figure 3.5. 5 – Histogram of Standardized Number of Frequencies Distribution / **Annex, Figures 3.5.1**

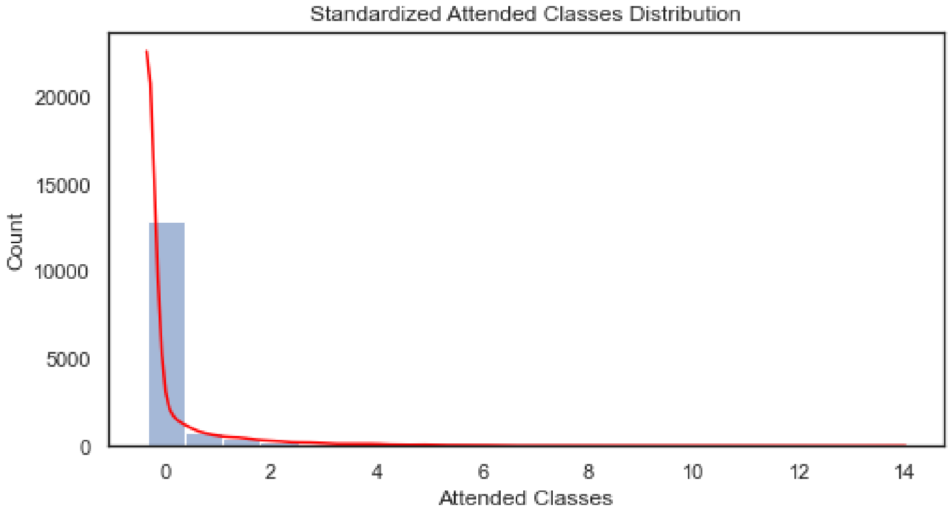


Figure 3.5. 6 – Histogram of Standardized Attended Classes Distribution / **Annex, Figures 3.5.1**

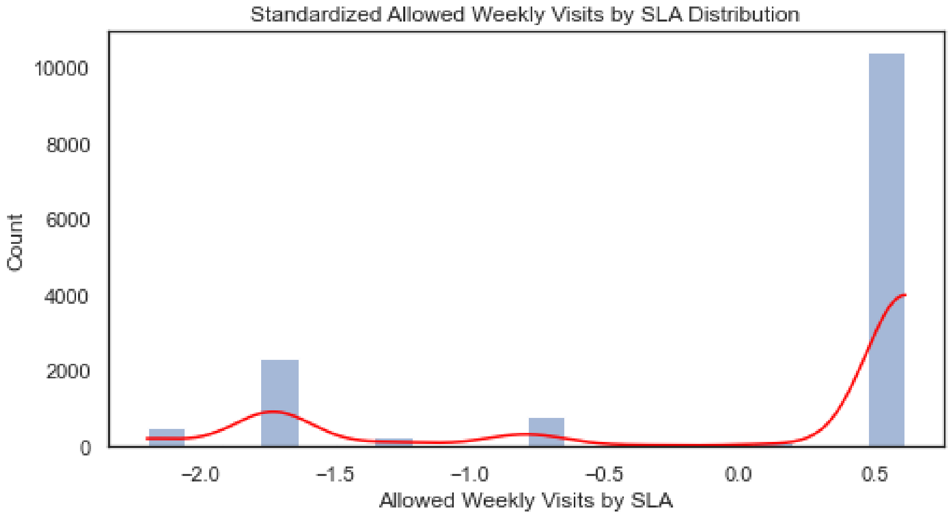


Figure 3.5. 7 – Histogram of Standardized Allowed Weekly Visits by SLA Distribution / **Annex, Figures 3.5.1**

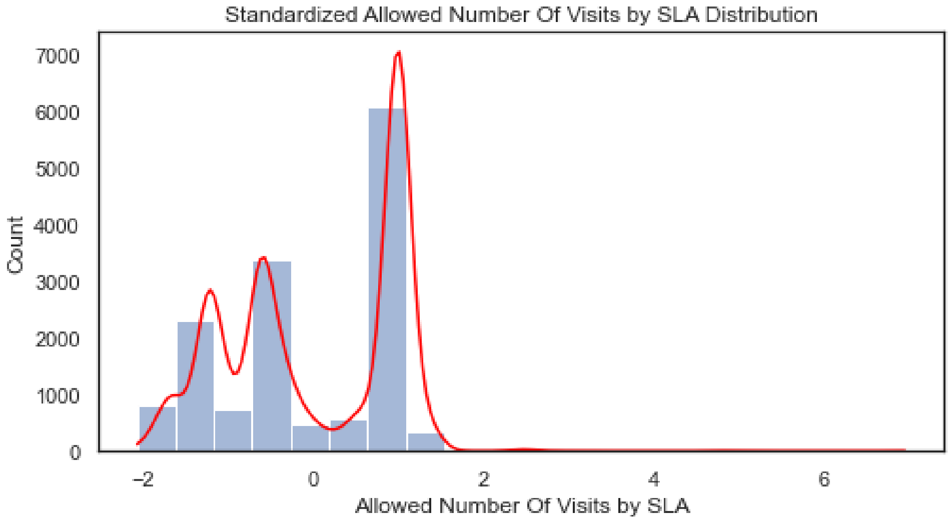


Figure 3.5. 8 – Histogram of Standardized Allowed Number Of Visits by SLA Distribution / **Annex, Figures 3.5.1**

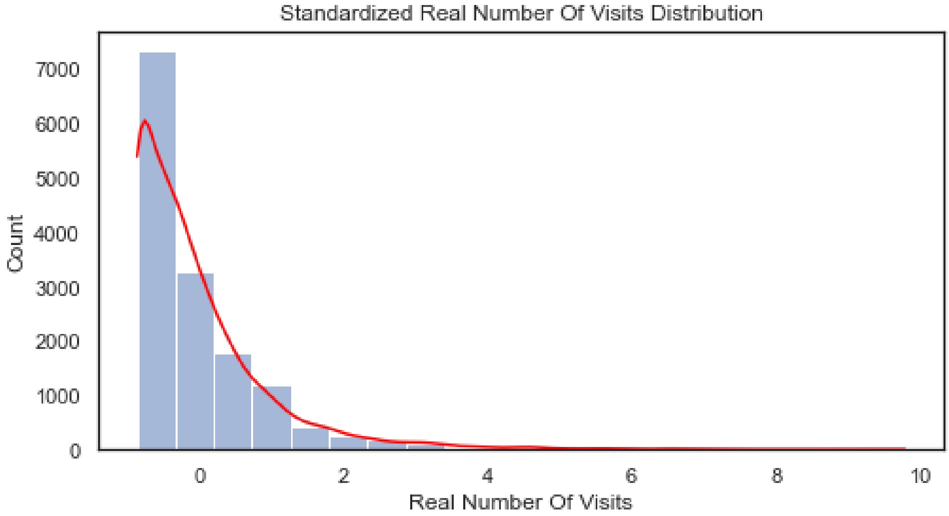


Figure 3.5. 9 – Histogram of Standardized Real Number Of Visits Distribution / **Annex, Figures 3.5.1**

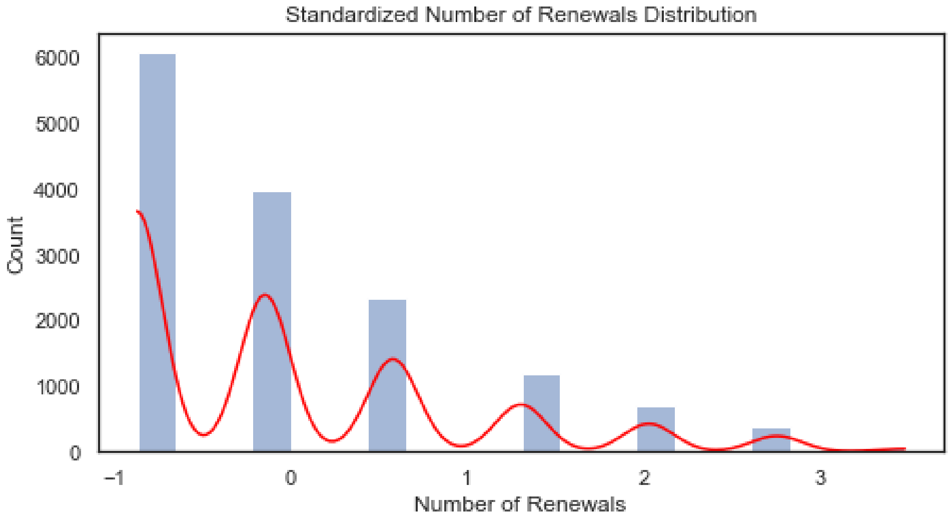


Figure 3.5. 10 – Histogram of Standardized Number of Renewals Distribution / **Annex, Figures 3.5.1**

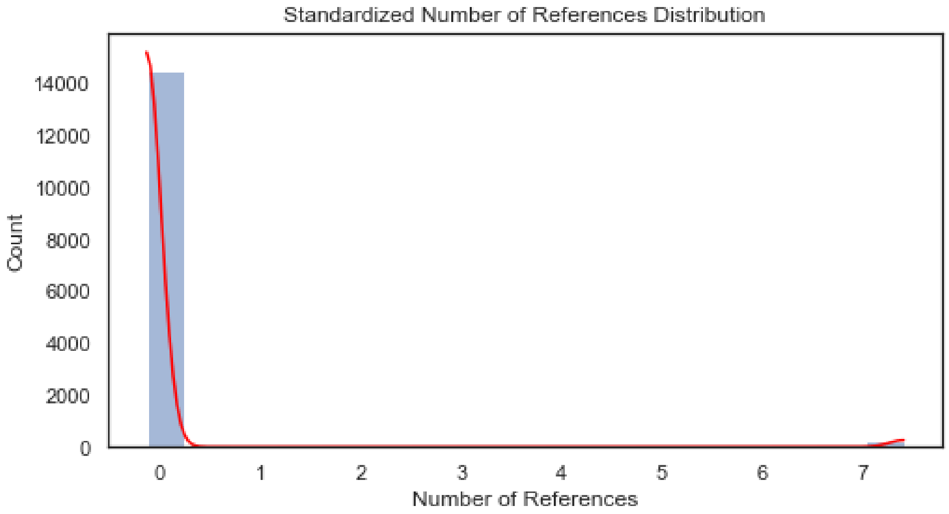


Figure 3.5. 11 – Histogram of Standardized Number of References Distribution / **Annex, Figures 3.5.1**

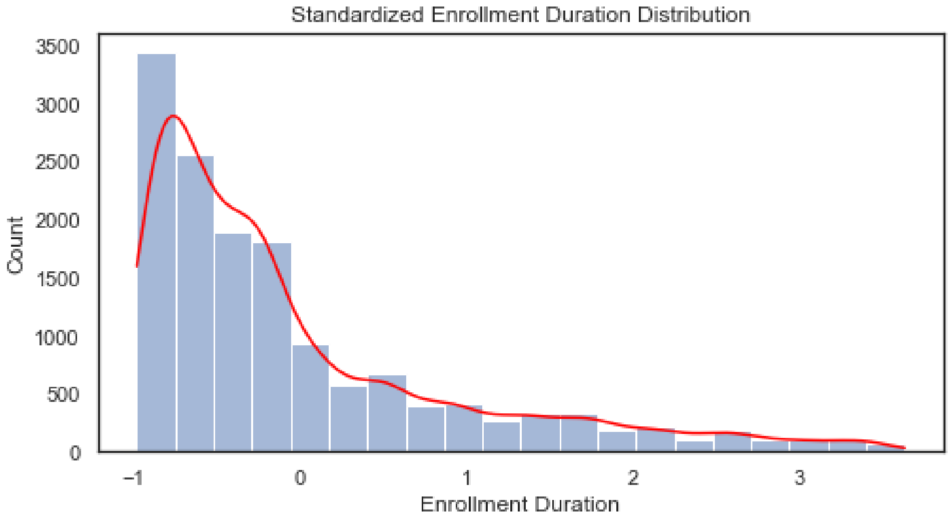


Figure 3.5. 12 – Histogram of Standardized Enrollment Duration Distribution / **Annex, Figures 3.5.1**

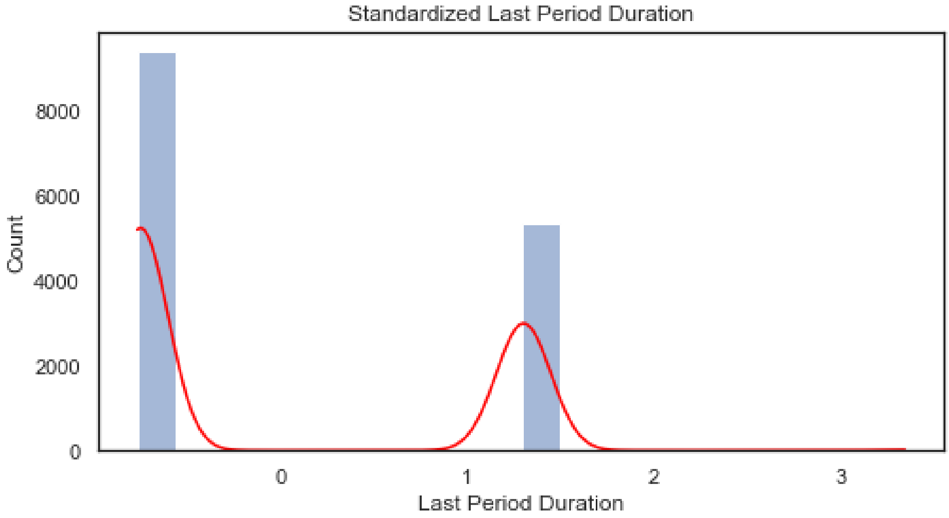


Figure 3.5. 13 – Histogram of Standardized Last Period Duration Distribution / **Annex, Figures 3.5.1**

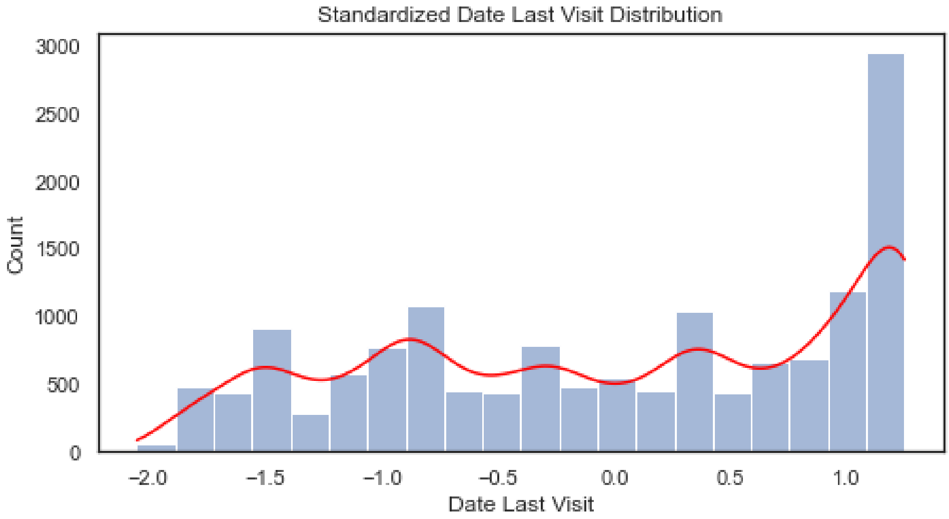


Figure 3.5. 14 – Histogram of Standardized Date Last Visit Distribution / **Annex, Figures 3.5.1**

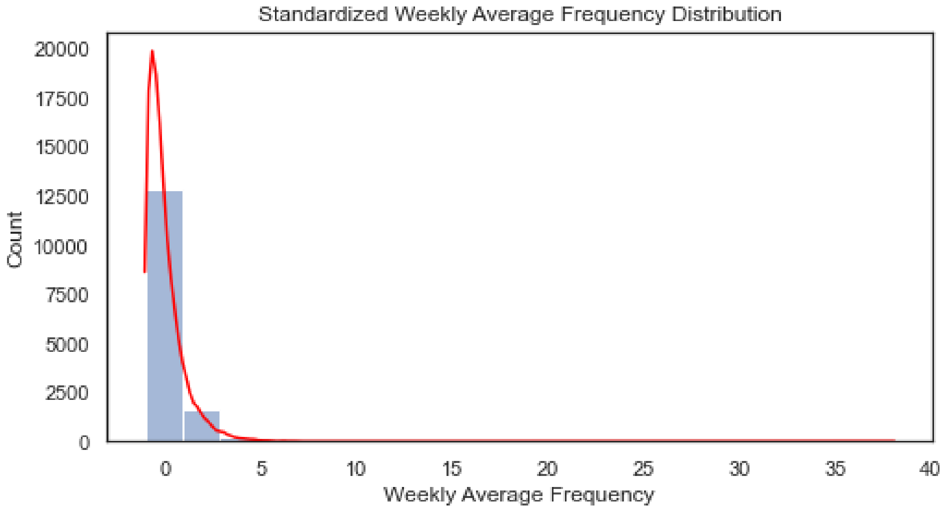


Figure 3.5. 15 – Histogram of Standardized Weekly Average Frequency Distribution / **Annex, Figures 3.5.15**

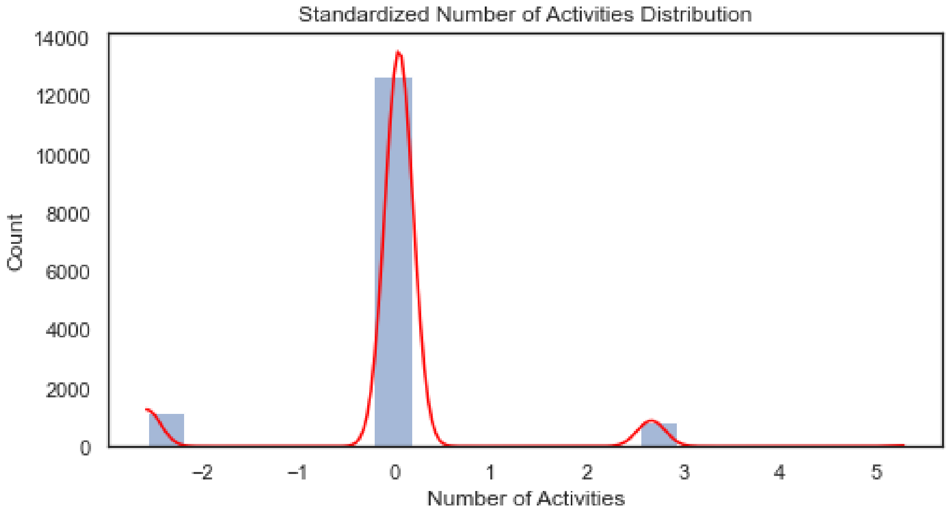


Figure 3.5. 16 – Histogram of Standardized Number of Activities Distribution / **Annex, Figures 3.5.16**

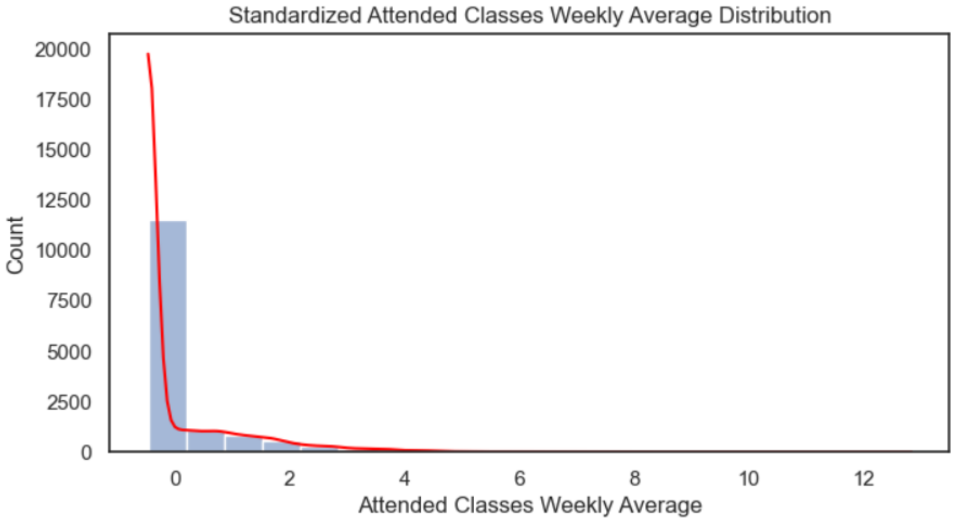


Figure 3.5. 17 – Histogram of Standardized Attended Classes Weekly Average Distribution / **Annex, Figures 3.5.17**

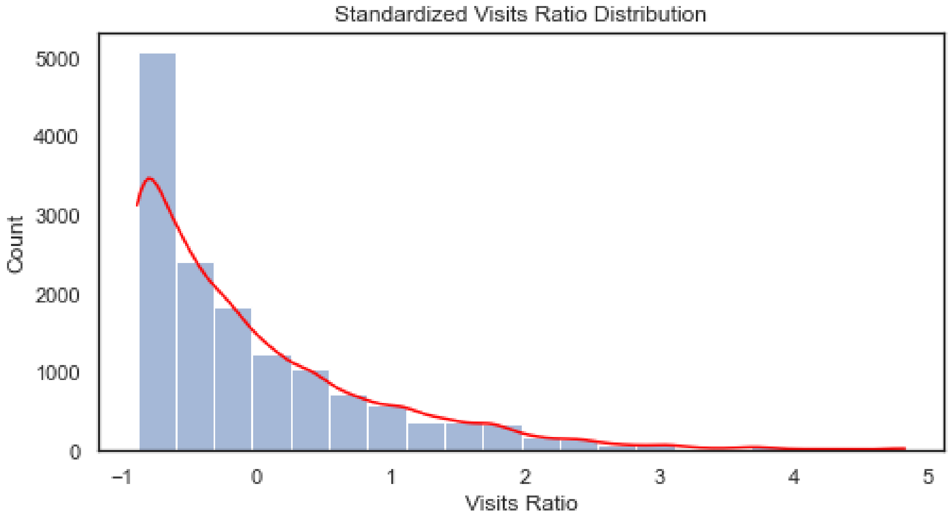
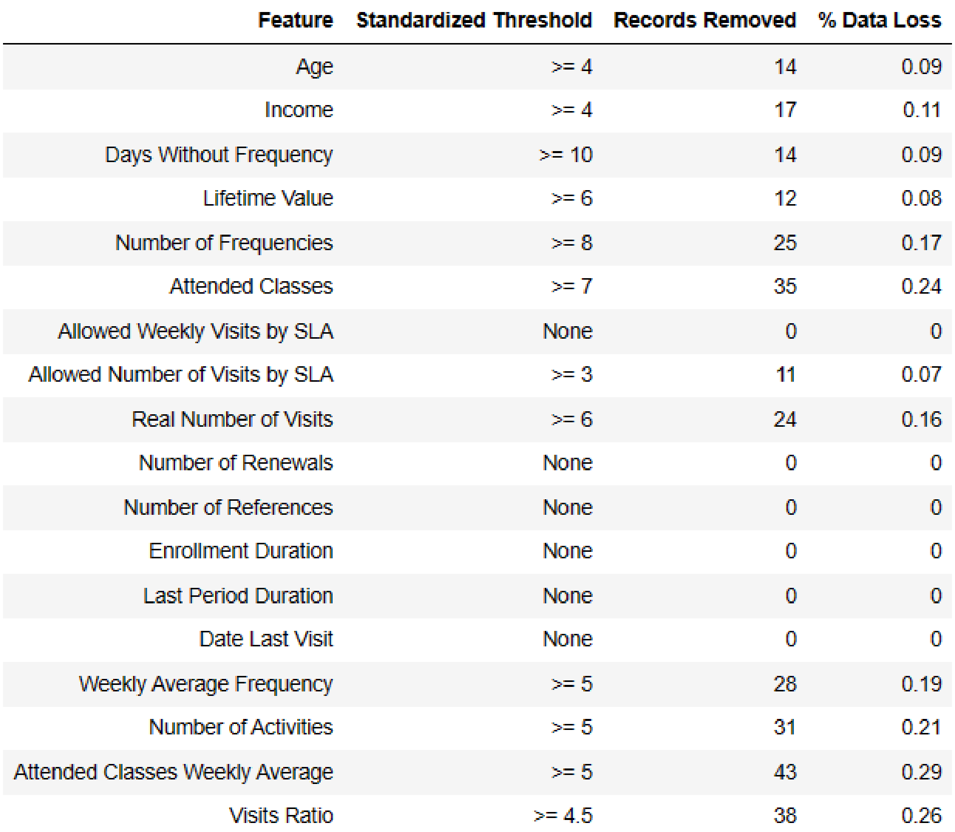


Figure 3.5. 18 – Histogram of Standardized Visits Ratio Distribution / **Annex 3.5.18.**



Feature	Standardized Threshold	Records Removed	% Data Loss
Age	>= 4	14	0.09
Income	>= 4	17	0.11
Days Without Frequency	>= 10	14	0.09
Lifetime Value	>= 6	12	0.08
Number of Frequencies	>= 8	25	0.17
Attended Classes	>= 7	35	0.24
Allowed Weekly Visits by SLA	None	0	0
Allowed Number of Visits by SLA	>= 3	11	0.07
Real Number of Visits	>= 6	24	0.16
Number of Renewals	None	0	0
Number of References	None	0	0
Enrollment Duration	None	0	0
Last Period Duration	None	0	0
Date Last Visit	None	0	0
Weekly Average Frequency	>= 5	28	0.19
Number of Activities	>= 5	31	0.21
Attended Classes Weekly Average	>= 5	43	0.29
Visits Ratio	>= 4.5	38	0.26

Figure 3.5. 19 – Metric Features Manual 2nd Outlier Removal / **Annex, Figure 3.5.19**

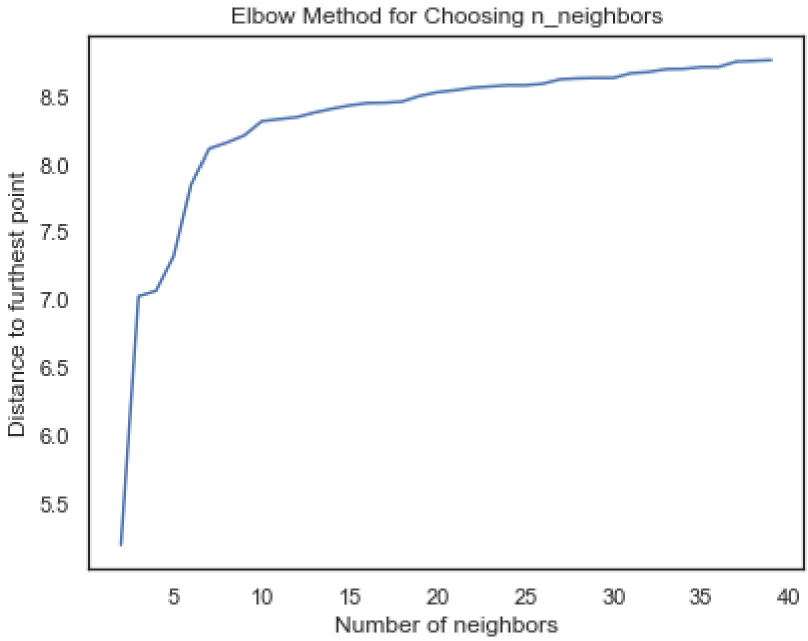


Figure 3.5. 20 – Elbow Method for choosing the number neighbors / **Annex, Figure 3.5.20**

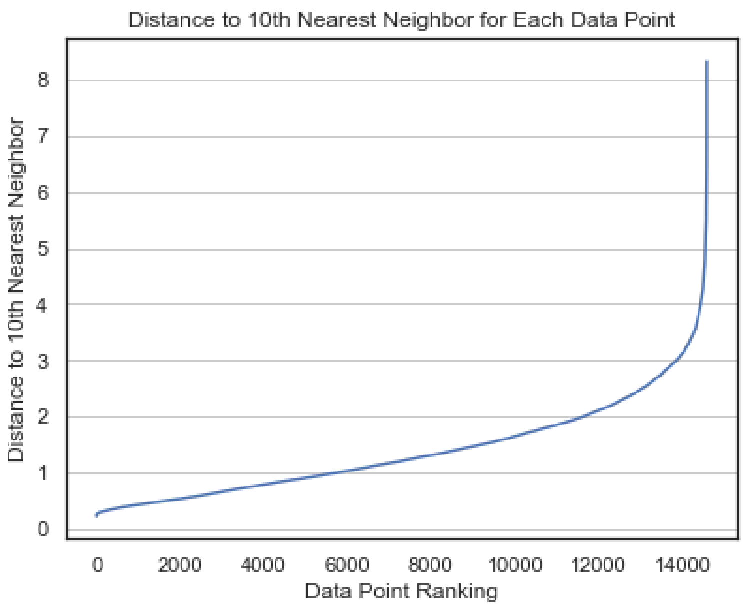


Figure 3.5. 21 – Distance to 10th Nearest Neighbor for Each Data Point / **Annex, Figure 3.5.21.**

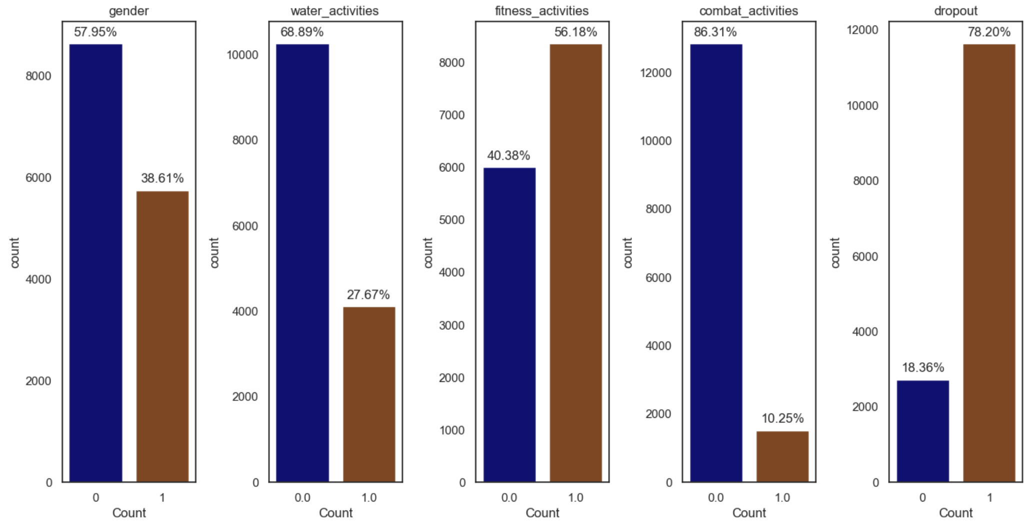


Figure 3.5. 22 – Binary Feature Distribution after 2nd Outlier Removal / **Annex Annex, Figure 3.5.22**

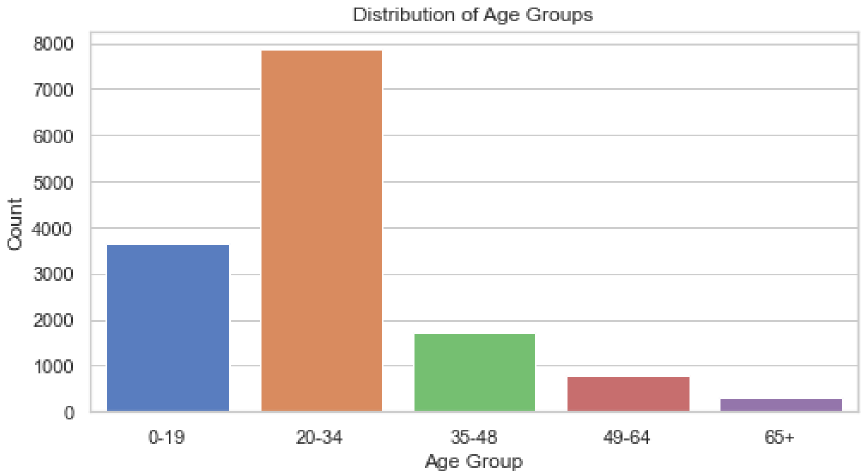


Figure 3.5. 23 – Bar plot of the distribution of age groups / **Annex, Figure 3.5.23.**

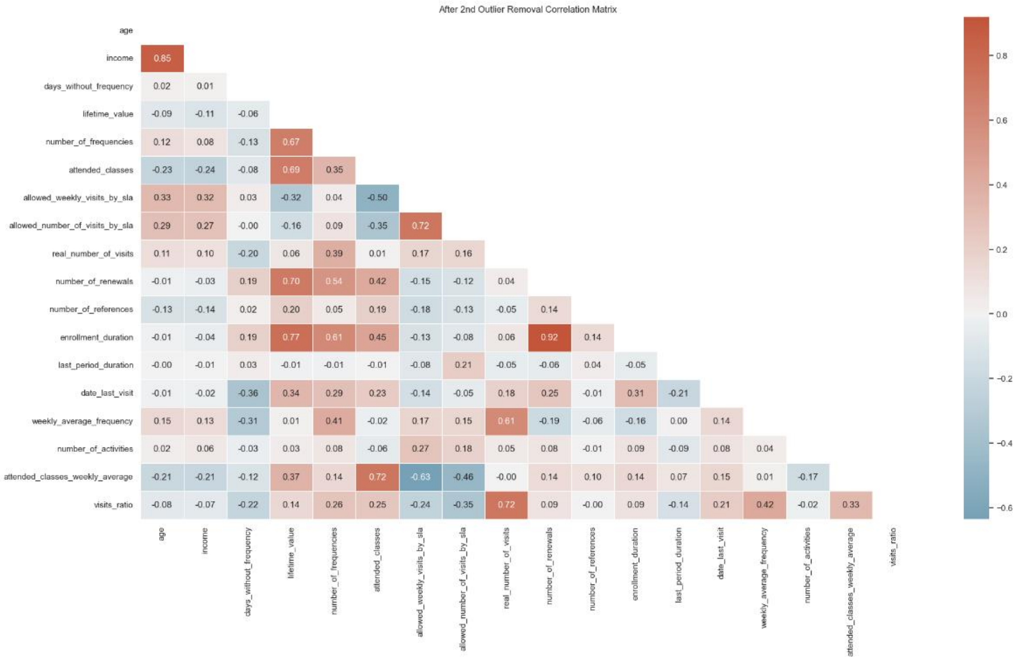


Figure 3.6. 1 – After 2nd Outlier Removal Correlation Matrix / **Annex, Figure 3.6.1.**

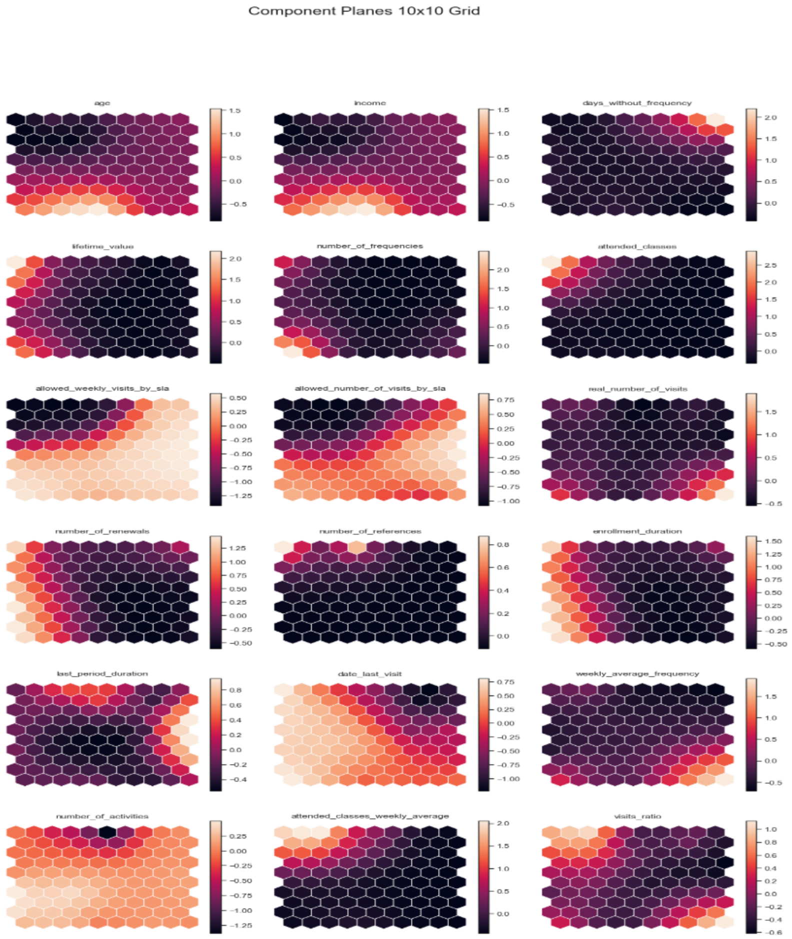


Figure 3.6. 2 – Feature Selection on Metric Features using Self Organizing Maps / **Annex, Figure 3.6.2.**

FIGURE 3.6. 3 – Explanation:

The Phi coefficient is a measure of association between two binary variables, similar to the Pearson correlation coefficient but designed for binary (0 or 1) data. It ranges from -1 (perfect negative association) to +1 (perfect positive association), with 0 indicating no association. The coefficient is calculated using a 2x2 contingency table of the variables' occurrences. It is important for measuring linear relationships in binary data, though it doesn't imply causation and is not suitable for non-binary categorical data.

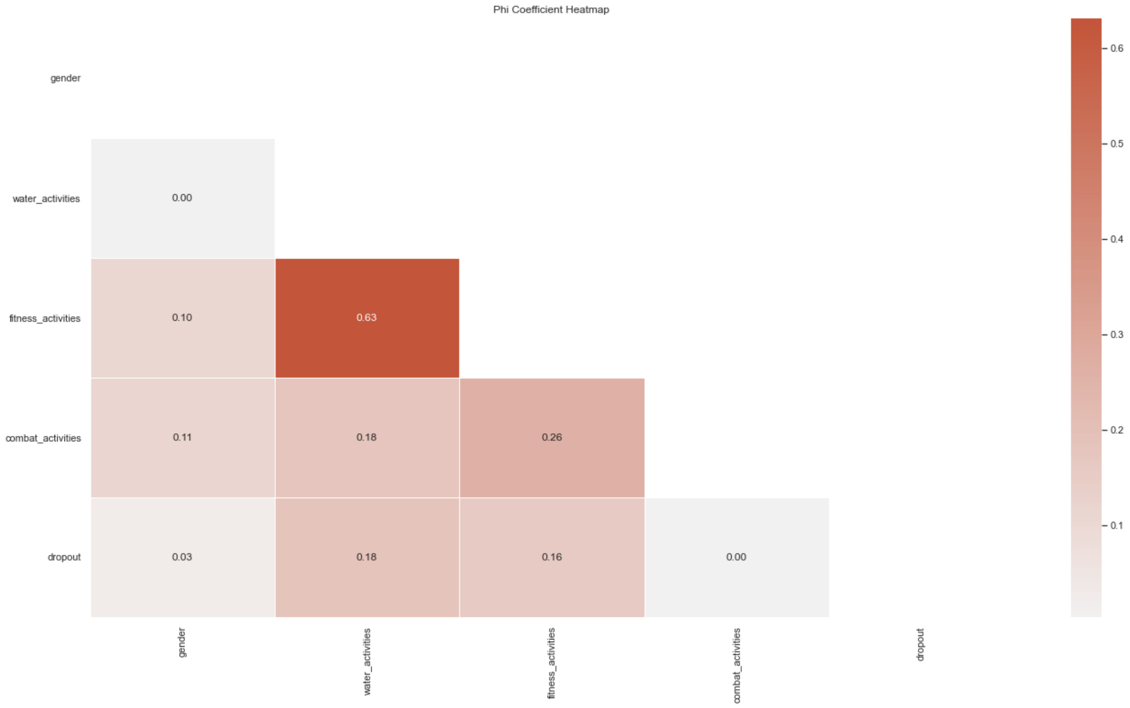


Figure 3.6. 4 – Feature Selection using Phi Coefficient Heatmap / **Annex 3.6.3.**

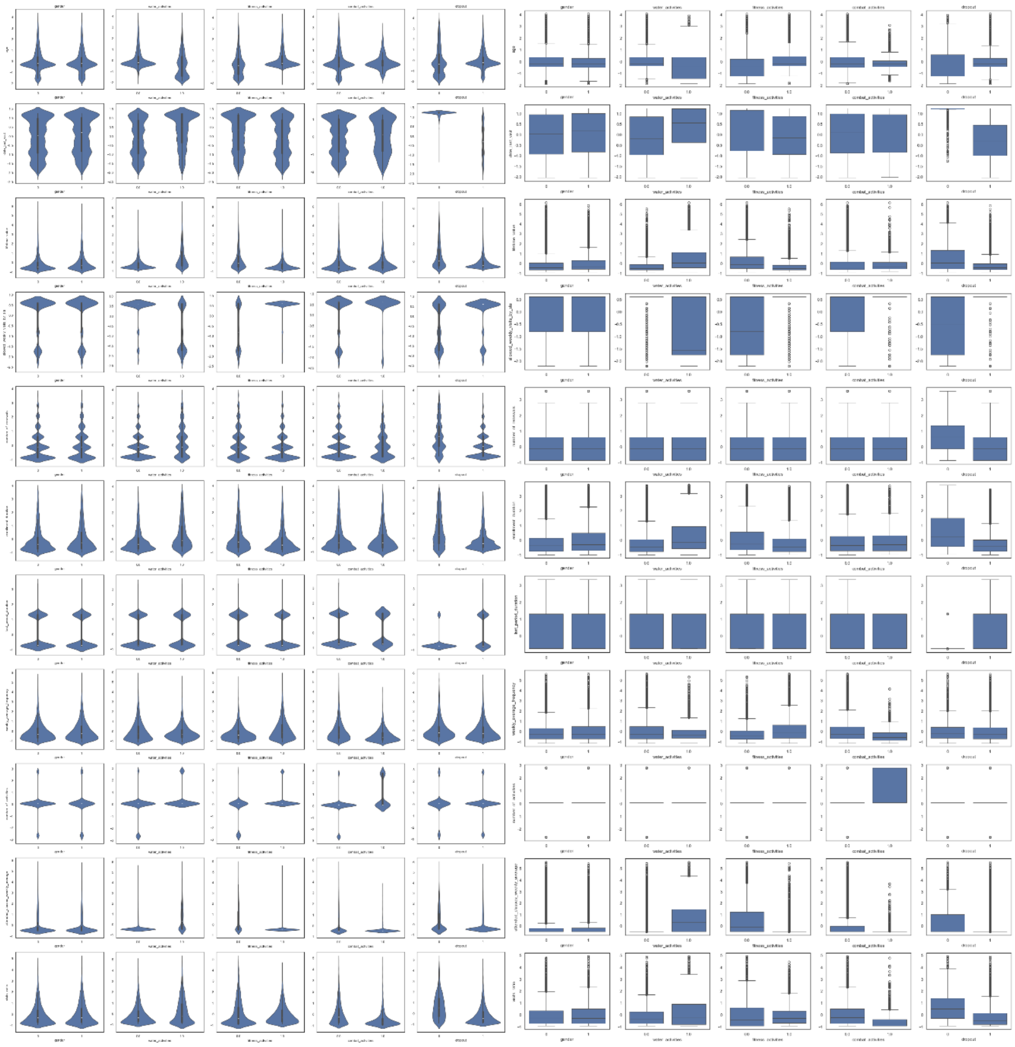


Figure 3.6. 5 – Feature Selection on Binary Features Boxplot & Violin plots Analysis / **Annex, Figure 3.6.4.**

FIGURE 3.6. 5 – Explanation:

The Point-Biserial Correlation is a statistical measure used to determine the relationship between a binary categorical variable and a continuous variable. It's a specialized form of the Pearson correlation coefficient, designed for cases where one variable is dichotomous (like 'yes' or 'no') and the other is continuous. The correlation coefficient ranges from -1 to +1, indicating perfect negative or positive relationships, respectively, or 0 for no significant association. It's useful for analyzing how a binary outcome relates to a continuous predictor, for example, in assessing how test scores correlate with pass/fail outcomes. However, it's important to note that this method only measures linear relationships and does not imply causation.

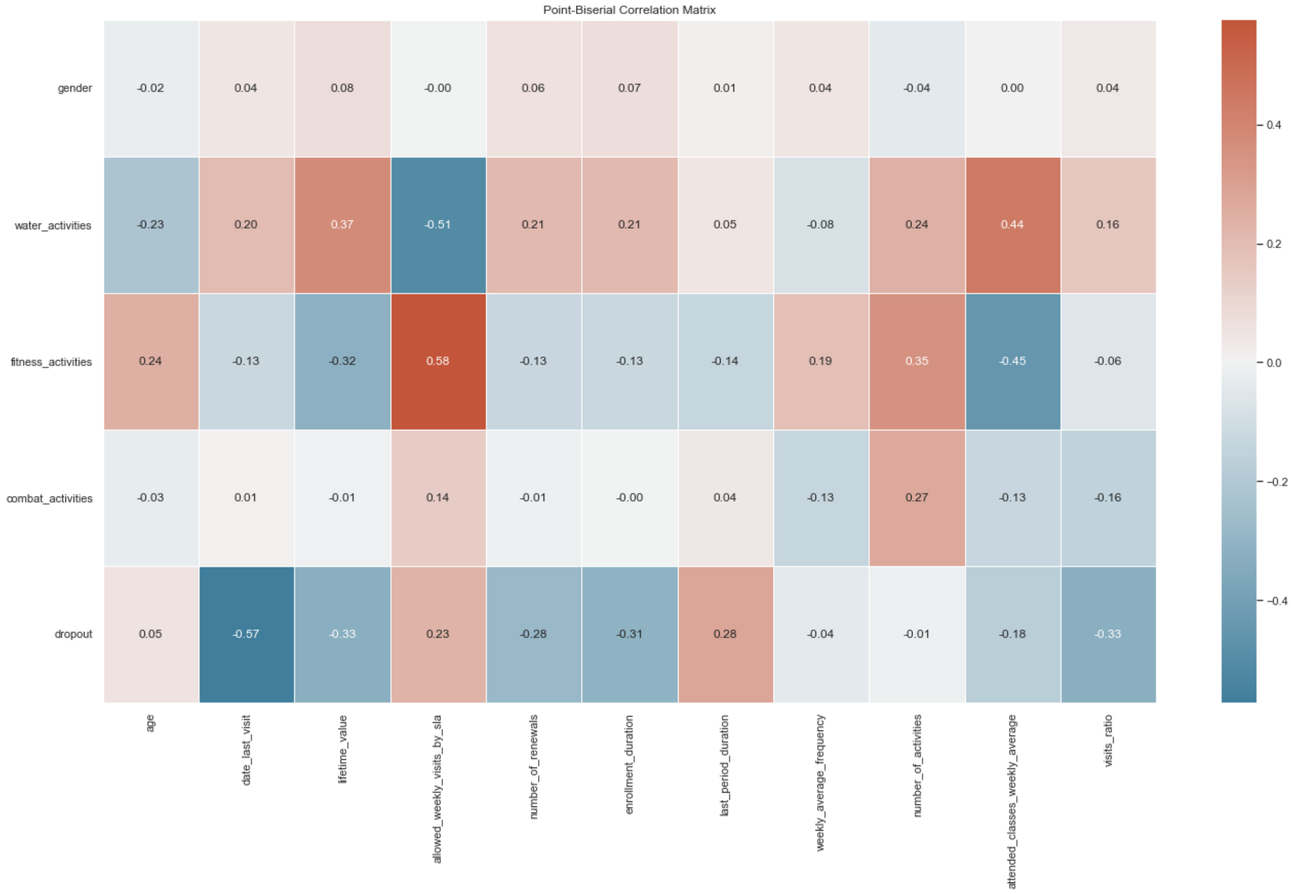


Figure 3.6. 6 – Feature Selection on Binary Features using Point-Biserial Correlation Matrix / **Annex 3.6.5**

The Calinski-Harabasz index is a statistical measure used to evaluate the quality of clustering in a dataset, particularly helpful for determining the optimal number of clusters. It calculates the ratio of the between-cluster variance (which assesses how distinct the clusters are from each other) to the within-cluster variance (which measures how compact the clusters are internally). A higher Calinski-Harabasz score indicates denser and better-separated clusters, implying a more effective clustering structure. This index is especially useful in methods like K-means for selecting the appropriate number of clusters by comparing the index values across different cluster counts. However, it has limitations, including its assumption of evenly distributed clusters and potential underperformance with non-spherical or varied-sized clusters.

Explanation 4.1 - Calinski-Harabasz Index in Detail / **Annex 4.1.**

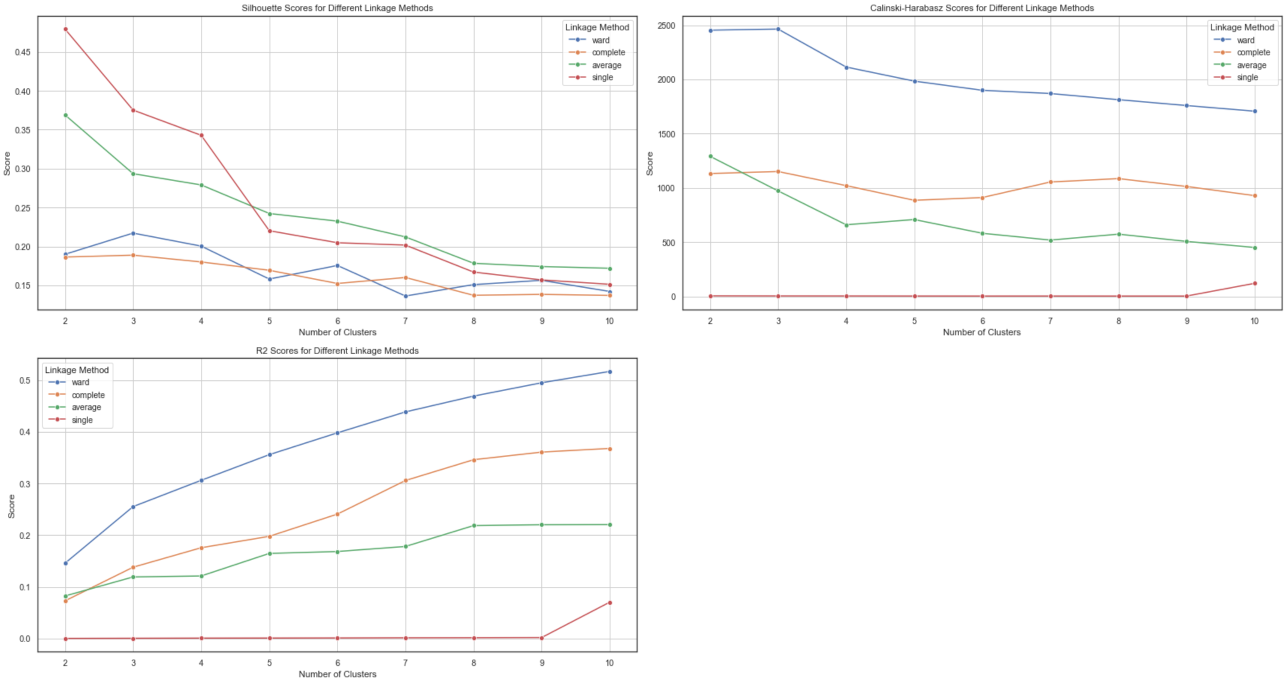


Figure 4.2 Hierarchical Agglomerative Clustering Evaluation Metrics / **Annex, Figure 4.2.**

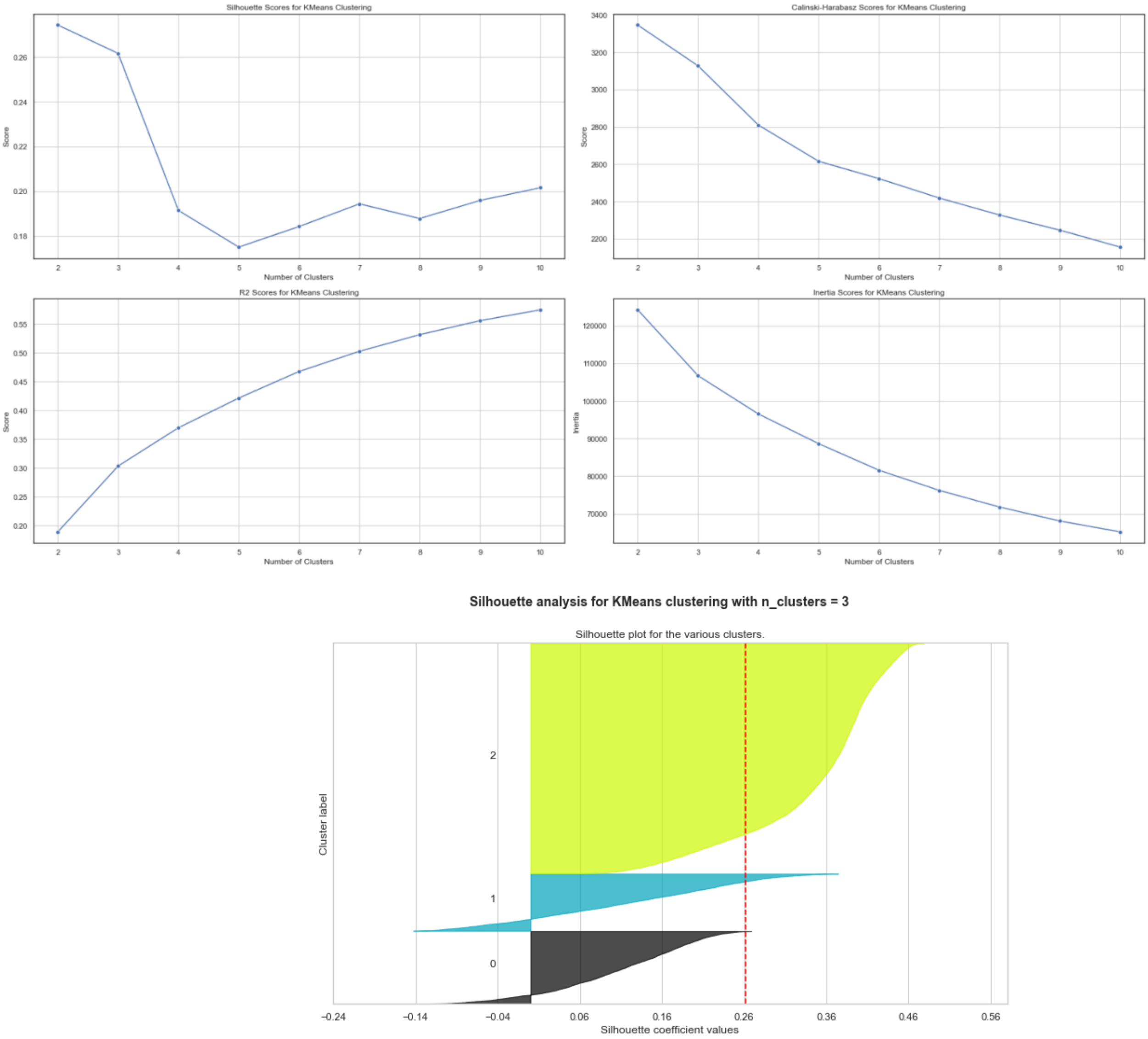


Figure 4.3 K-Means Clustering Evaluation Metrics / **Annex, Figure 4.3.**

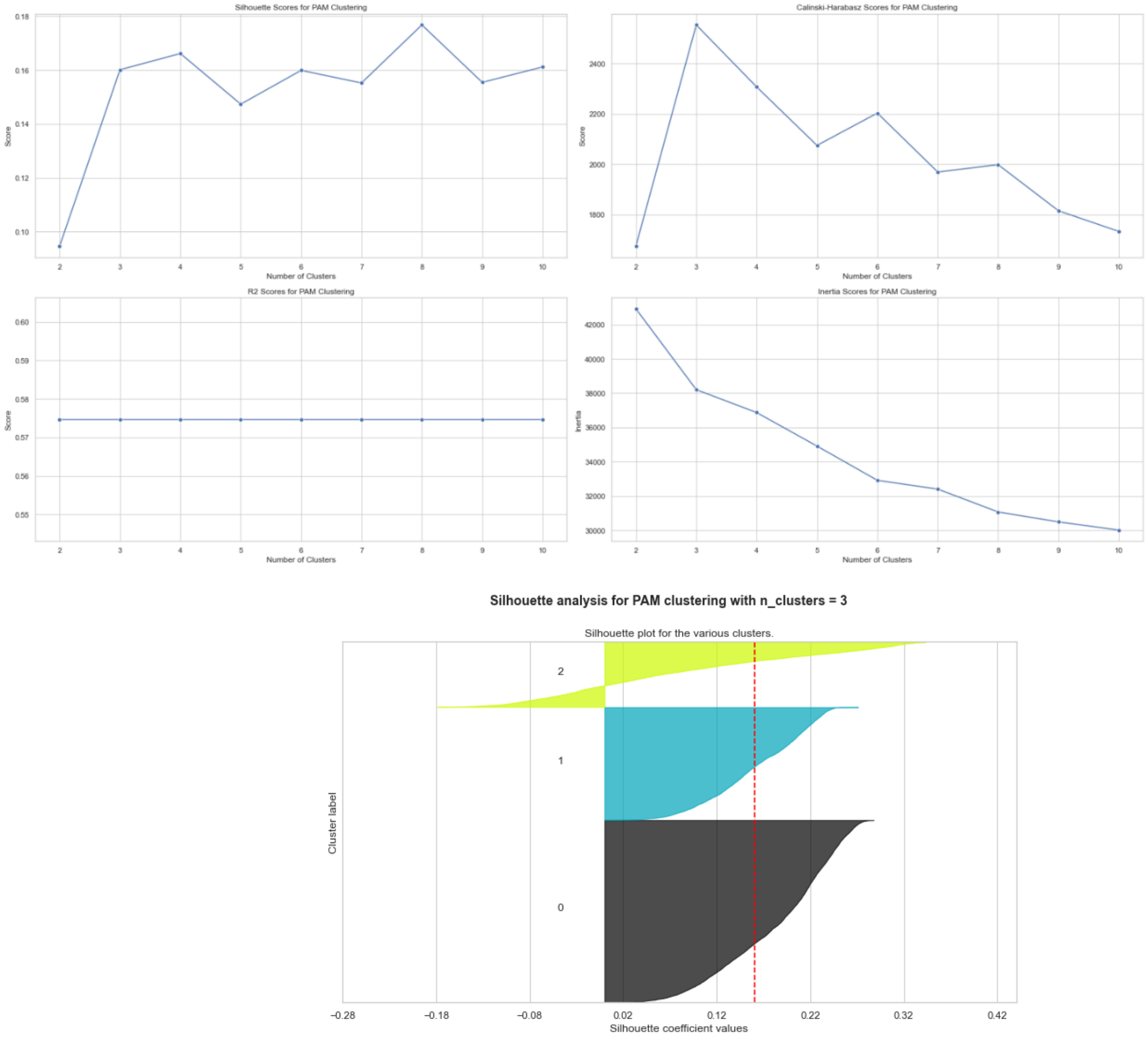


Figure 4.4 PAM Clustering Evaluation Metrics / **Annex, Figure 4.4.**

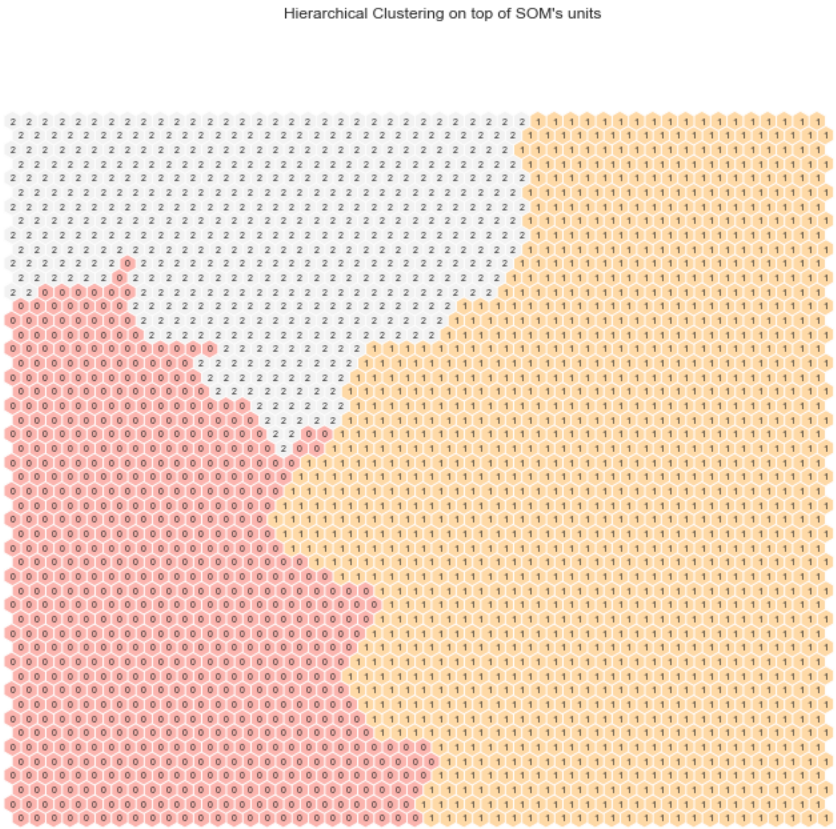


Figure 4.5. 1 – Hierarchical Clustering on top of Self Organizing Maps' units / **Annex 4.5.1**

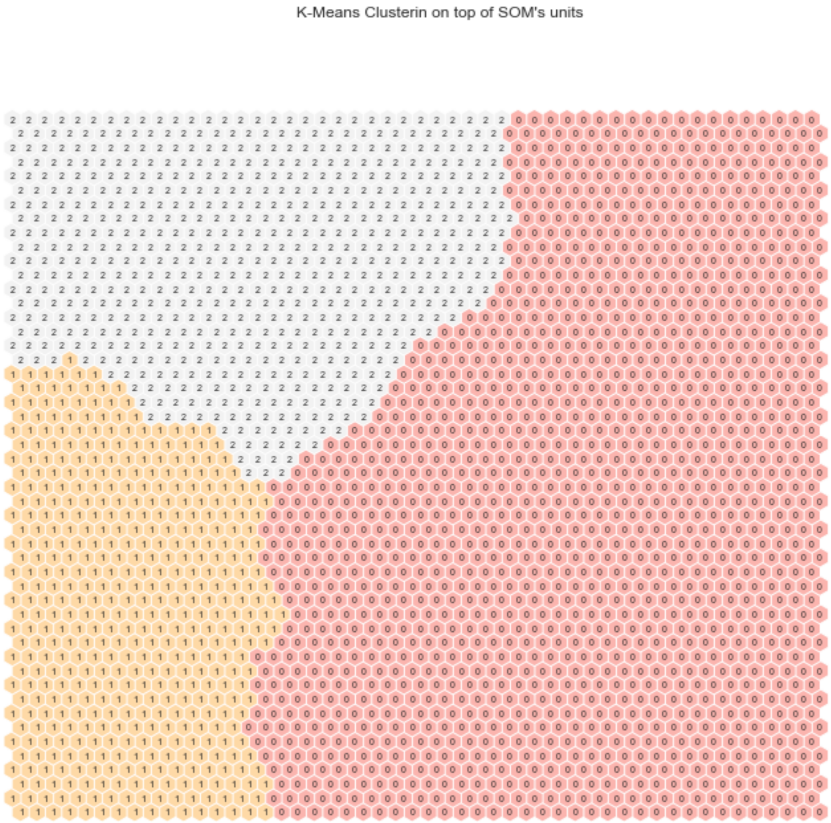


Figure 4.5. 2 – K-Means on top of Self Organizing Maps' units / **Annex 4.5.2**

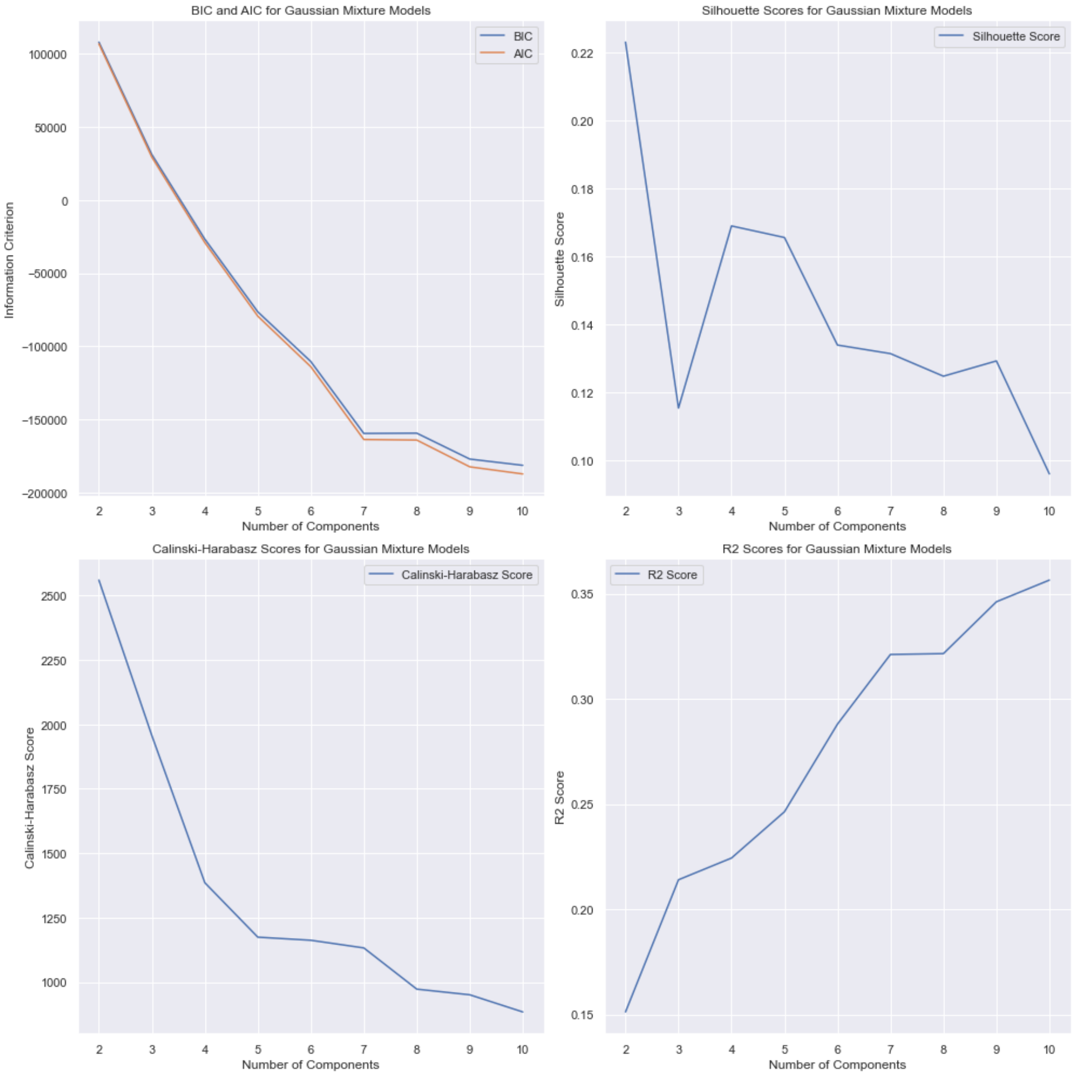


Figure 4.6. 1 – GMM Clustering Evaluation Metrics / **Annex 4.6.1**

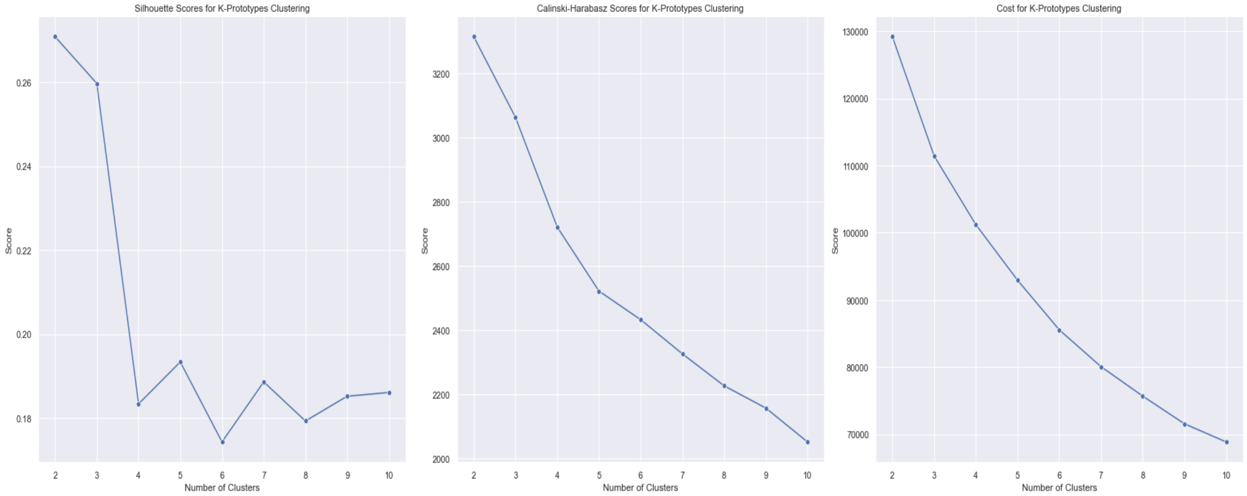
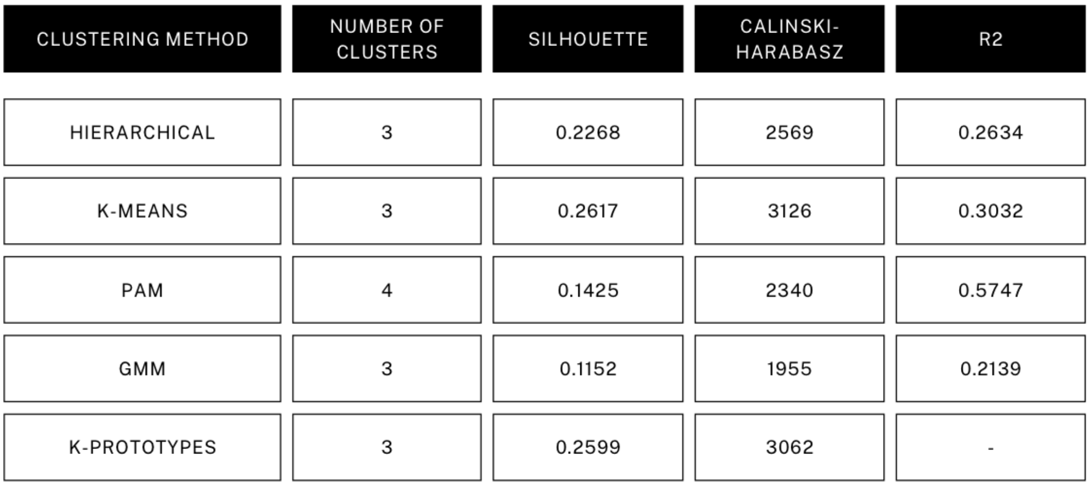


Figure 4.7. 1 K-Prototypes Clustering Evaluation Metrics / **Annex, Figure 4.7.1.**



CLUSTERING METHOD	NUMBER OF CLUSTERS	SILHOUETTE	CALINSKI-HARABASZ	R2
HIERARCHICAL	3	0.2268	2569	0.2634
K-MEANS	3	0.2617	3126	0.3032
PAM	4	0.1425	2340	0.5747
GMM	3	0.1152	1955	0.2139
K-PROTOTYPES	3	0.2599	3062	-

Figure 4.8 Clustering Methods Wrap-Up / **Annex, Figure 4.8.**

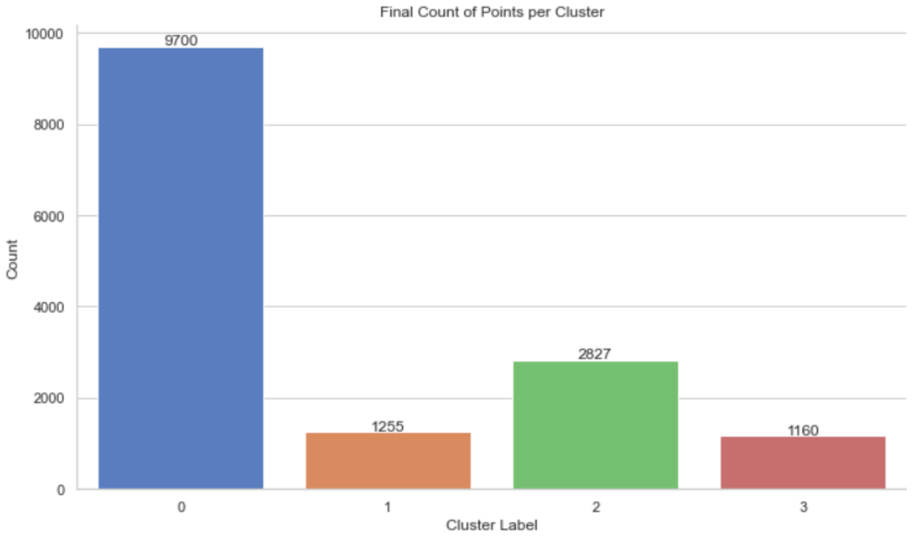


Figure 5.1 1 Final Count of Points Per Cluster / **Annex, Figure 5.1.1**

BIBLIOGRAPHY

[1] P. Pinheiro and L. Cavique, “Regular sports services: Dataset of demographic, frequency and service level agreement,” *Data in Brief*, vol. 36, p. 107054, Jun. 2021, doi: 10.1016/j.dib.2021.107054.

[2] “Phi Coefficient (Mean Square Contingency Coefficient) - Statistics How To.” Accessed: Jan. 06, 2024. [Online]. Available: https://www.statisticshowto.com/phi-coefficient-mean-square-contingency-coefficient/

[3] “t-Test, Chi-Square, ANOVA, Regression, Correlation...” Accessed: Jan. 06, 2024. [Online]. Available: https://datatab.net/tutorial/point-biserial-correlation

[4] M. Botyarov and E. E. Miller, “Partitioning around medoids as a systematic approach to generative design solution space reduction,” *Results in Engineering*, vol. 15, p. 100544, Sep. 2022, doi: 10.1016/j.rineng.2022.100544.

[5] Y. Reddy, “K-means, kmodes, and k-prototype,” Medium. Accessed: Jan. 06, 2024. [Online]. Available: https://medium.com/@reddyyashu20/k-means-kmodes-and-k-prototype-76537d84a669